Subscription Rates
(Effective from 1996)

	Institutional	Individuals
All countries except India (Price includes AIR MAIL charges)	US$100	US$40
India	Rs. 150	Rs. 75

All correspondence regarding subscription should be addressed to **The Circulation Department**
of the Academy.

Editorial Office

Indian Academy of Sciences, C V Raman Avenue
P.B. No. 8005, Sadashivanagar
Bangalore 560 080, India

Telephone: (080) 3342546, 3342943
Telefax: 91-80-3346094

e-mail: sadhana@ias.ernet.in

"Notes on the preparation of papers" are printed in the last issue of every volume.

Sādhanā

Academy Proceedings in Engineering Sciences

Volume 22, 1997

CONTENTS

i

Competitive manufacturing systems

Foreword

This is the first of two special issues on the important topic of competitive manufacturing systems. This topic is motivated by the heightened importance of the manufacturing domain in the wake of intense global competition and tremendous advances in technology. So, when Professor Viswanadham suggested in March 1996 that we bring out a Special Issue of *Sādhanā* on this topic, we readily agreed.

Competitive Manufacturing Systems

A competitive manufacturing system designs, produces, and delivers high-quality, customer-desired products faster than competition, in a dynamic and uncertain environment. World-wide, competition is now very intense and customers have exacting demands of low cost, low defect rates, high performance, and easy maintenance. Product life cycles are shrinking rapidly and new technology innovations are bringing about dramatic changes in the manufacturing domain. The basis for competing with other challengers has gradually shifted from cost and quality to time and core-competence. A world-class manufacturing enterprise should be able to perform at a high level with agility in the face of such dynamic and uncertain factors sweeping the manufacturing world. Researchers and practitioners have identified several core issues for the effective design, management, and continuous improvement of such high-performance manufacturing systems. These include: Flexible automation, business process management, flexibility management, systems integration and strategic networking.

Scope of the Special Issue

Since the success of a manufacturing enterprise depends not merely on the effectiveness of the factory floor operations but more critically on elements such as suppliers, logistics network, distributors, and new product designs, the scope of the special issue encompasses all these elements of a manufacturing enterprise. Also, the traditional "functional" view of an enterprise as a sequential arrangement of functions such as marketing, finance, R & D, and manufacturing is giving way to a more logical "process" view that regards an enterprise as a conglomerate of value-delivering business processes. These processes cut across multiple functional domains and emphasize how the organization delivers value to the customers. Important business processes include: Factory floor process, new product development process, plant maintenance process, supply chain process, and order-to-delivery process. The scope of this special issue includes business process modelling, analysis, and

management. In addition to the above "cosmic" and "process" view of a manufacturing organization, the special issue also focuses on important traditional topics such as quality, flexibility, modelling and simulation, scheduling etc.

Overview of the First Special Issue

We now give an overview of the seven papers appearing in this, the first special issue.

The first paper is by Mathur & Pattipati and is on the topic of off-line quality engineering methods, which have aroused widespread interest following Taguchi's robust design methodology. Quality engineering is an important methodology for controlling, maintaining, and improving product quality. It addresses two main types of problems: On-line quality control and off-line design of products and processes for quality. In this paper, Mathur & Pattipati provide an excellent overview of the evolution of off-line quality engineering methods with single or multiple quality criteria. Since most design-for-quality problems involve multiple quality criteria, the authors emphasize various methods available for multi-objective optimization techniques for robust parameter design. This article is rich with examples and provides a clear overview of the problems, methodologies, and recent results.

In the second paper, authors Ramesh Srinivasan & Jayashankar Swaminathan treat an important problem of planning and coordination that results when manufacturers strive to offer broader product lines to achieve greater customer satisfaction. Focusing on the computer industry, the authors study the increase in the complexity of the manufacturing process and in the complexity of forecasting, parts planning, final assembly, and delivery of products entailed by the effort to offer more variety in product lines. In particular, the authors consider a feature-based product line in the computer industry and highlight issues related to forecasting, parts planning, final assembly, and inter-plant coordination. Techniques adopted to overcome the challenges are also described.

The third paper, by Sridhar Tayur, provides an overview of the recent developments in the single product, discrete-time, capacitated production-inventory models. The model is motivated by typical supply chain design problem in a diverse set of companies. The author brings out the interaction between demand variability, available production capacity, inventory holding costs, lead times, and desired service levels. Several significant advances have occurred in the study of such models in the last few years and the author has provided a clear overview of the problems and recent techniques.

The next two papers are on scheduling. The first one is concerned with stochastic scheduling with setups in a single flexible machine, and the second one with the heuristic, deterministic scheduling of an assembly job shop. Narahari & Hemachandra focus on a flexible machine serving multiple classes of products, with a setup or switch-over time and/or cost involved whenever the machine switches to a different product type. Choosing the total cost as the sum of discounted inventory and setup costs over an infinite horizon, they prove the optimality of exhaustive service policies for a fairly general arrival pattern of jobs. Their result generalizes the earlier work of many researchers who were constrained to assume independent Poisson inputs. The arrivals considered by Narahari & Hemachandra are more general; in particular, correlated in the Markov modulated sense. The second paper on scheduling is by Awate & Saraph. This paper addresses the issue of scheduling

assembly job shops in contrast to the more popular job-shop scheduling. It advocates
 interesting look-back approach to take into account the staging delays suffered by
perationally late jobs. Their scheduling method generalizes several existing approaches
d outperforms well-known heuristics.

The sixth paper in this issue is by Raviraju, Reddy & Krishnaiah Chetty. This paper
ddresses an important class of modern manufacturing systems, namely Just-in-Time pro-
uction systems, and makes a case for modelling them using priority nets. The priority net
odels are simulated to gain insights into many performance issues and the sensitivity of
erformance to important input parameters.

Developments in software technology have been at the heart of many advances in the
anufacturing domain. Manufacturing planning and control software belongs to the realm
 software in the large. The development of large-scale and complex software products,
ch as in the manufacturing world, is a complex and expensive process. Sarma & Vijay
ao, in the last paper of this issue, present queueing models for the software testing
ocess in such a software product development environment. The models are based on
e re-entrant line model, which itself was proposed in the first place to model re-entrant
anufacturing systems such as semiconductor fabs and thin film lines.

We would like to acknowledge the continuous support and ideas from Professor
iswanadham. The authors have been most co-operative and have taken minimal time with
eir papers. All the reviewers have been very prompt and thorough in their reviews. It is
ostly because of this wonderful, quick response from the authors and the reviewers that
is special issue has seen the light in exactly one year's time (from conception to market).
e Academy staff, coordinated efficiently by Ms. Shashikala, have also contributed sig-
ficantly to the compression of this issue's lead time. We would also like to acknowledge
e support provided by our respective departments: Computer Science and Automation
 the Indian Institute of Science; School of Industrial Engineering and Management at
e Oklahoma State University; and the Kellogg Graduate School of Management at the
orthwestern University.

bruary 1997

<div align="right">

Y NARAHARI
M KAMATH
RAVI ANUPINDI
Guest Editors
</div>

Single- and multiobjective optimization problems in robust parameter design

AMIT MATHUR and KRISHNA R PATTIPATI

Department of Electrical and Systems Engineering, University of Connecticut, Storrs, CT 06269, USA
e-mail: [amit,krishna]@sol.uconn.edu; amit@teamqsi.com

Abstract. This paper reviews the evolution of off-line quality engineering methods with respect to one or more quality criteria, and presents some recent results. The fundamental premises that justify the use of robust product/process design are established with an illustrative example. The use of designed experiments to model quality criteria and their optimization is briefly reviewed. The fact that most design-for-quality problems involve multiple quality criteria motivates the development of multiobjective optimization techniques for robust parameter design. Two situations are considered: one in which response surface models for the quality characteristics can be obtained using regression and considered over a continuous factor space, and one in which the problem scenario and the experiment permit only discrete parameter settings for the design factors. In the former scenario, a multiobjective optimization technique based on the reference-point method is presented; this technique also incorporates an inference mechanism to deal with uncertainty in the response surface models caused by finite, noisy data. In the discrete-factors scenario, an efficient method to reduce computational complexity for a class of models is presented.

Keywords. Multiobjective optimization; off-line quality engineering; Taguchi methods; robust product/process design.

1. Introduction

The primary goal of a quality program in the manufacturing industry is to design and implement a production process geared towards satisfying the customer. It concerns itself with all aspects of the process, from conceptualization through design, to production and operations. Besides the need for an environment of *total quality management* built on a concurrent engineering foundation, the *quality engineering* tools needed to evaluate and optimize physical quality characteristics are paramount to the success of a quality program. Development of some of these tools is the focus of this article.

Quality engineering deals with a methodology for maintaining (controlling) and [im]proving product quality. It addresses two main types of problems: on-line quality-co[ntrol] and off-line design of products and processes for quality. The on-line techniques have [gen]erally been researched and applied under the rubric of statistical process control (S[PC]) where the popular tools are various types of control charts (Montgomery 1991a). The [off-] line design-for-quality problem has aroused widespread interest in the West only sinc[e the] eighties with the popularization of Taguchi's methods of parameter design using desi[gn of] experiments (also known as *robust design*, Phadke 1989). While SPC aims at *maint[aining]* the quality of a production process at prescribed levels with the tacit assumption tha[t the] design has been optimized prior to production and that no further design improveme[nt is] necessary, off-line product development uses the tools of experimental design to o[btain] models for the various quality metrics of interest, which are then optimized with re[spect] to the design variables (factors). In both exercises, however, the ultimate objective [is to] ensure that the production system performs as close to the targeted levels and with as [little] variability as possible.

The prerequisites for a robust product/process-design exercise are the identific[ation] of suitable metrics or performance measures that can summarize quality, and the sp[ecifi]cation of targets for these measures. The objective then is to obtain values of the va[rious] product and process parameters that bring these measures to their targets. Manuf[actur]ing processes almost always require consideration of several individual quality cri[teria] whose physical relationships to the control parameters or design variables, as well a[s to] each other, are not always known. The problem, therefore, involves modelling all q[uality] characteristics with respect to the control parameters (input variables) and simult[aneous] optimization of these quality criteria. Since the same settings of control parameters, in [gen]eral, cannot optimize all objectives, a trade-off or compromise is inevitable. The f[ormal] methodology for doing so, viz., *multiobjective optimization* (also termed multiple-cr[iteria] decision-making or multiobjective programming), has been a subject of extensive res[earch] in the area of resource management and planning for more than two decades (Zeleny [1982; Chankong & Haimes 1983; Steuer 1989). On the other hand, quality engineering has [con]cerned itself mainly with the statistical and modelling aspects of the problem wit[h the] underlying expectation that the number of optimization variables would be small an[d that] the concomitant optimization problem would be solvable via *ad hoc* methods. Wit[h the] need for increasing sophistication in product design and process improvement in mo[dern] manufacturing, it is important to incorporate a formal methodology in quality engine[ering] that integrates both statistical modelling and optimization into a single framework. [This] is especially so as the number of quality criteria and the number of variables affectin[g the] process or product performance increase, rendering visual/graphical tools cumber[some] or ineffective. Multiobjective optimization methods can be applied to quality engine[ering] problems with only a few conceptual extensions.

This paper reviews the evolution of off-line quality engineering methods for s[ingle] and multiple objectives, and introduces some recent results. Since the focus is o[n the] engineering tools (also known as the CAE (computer-aided engineering) tools, Boza [et al] 1994), the issue of how to identify appropriate quality characteristics and the facto[rs of] interest in a production system is ignored – those objectives are best accomplishe[d by] extensive brainstorming by the quality team (Bendell *et al* 1989; Boza *et al* 1994). I[n the]

remainder of this article, we first give a general overview of experimental design techniques and the notion of robust design attributed to Taguchi. The optimization of quality metrics with respect to continuous factors is addressed first and then the optimization problem with respect to discrete factors is considered. For the case of continuous design factors, we also examine the optimization problem for the single as well as the multiple objective cases in presence of uncertainty in models estimated from experimental data. We present a strategy for design improvements that account for the uncertainty in regression models. All concepts and methods are illustrated by examples.

2. Overview of off-line design

2.1 *Quality metrics and experimental design*

The goal of Taguchi's design-for-quality approach is to choose levels (or settings) of factors (or design parameters) that minimize the variability of a product's function while achieving the target (Taguchi 1987; Phadke 1989; Nair 1992). This is in contrast with the traditional (and uneconomical) approach of 'accept–reject,' where products not satisfying the specifications within a tolerance are rejected. Taguchi's philosophy is based on the premise that quality can be built into a product by a proper choice of design settings to lower the inherent variability, and thus to lower the cost of producing quality products. This cost first needs to be quantified by a *quality loss function*. A simple quality loss function for a quantitative quality characteristic (or response) y, which is targeted to have a value T, is the quadratic loss function

$$L(y) = k_0(y - T)^2.$$

Figure 1 plots the quadratic loss function in relation to the step loss function that corresponds to the tolerance-based approach; δ is the tolerance limit about the target T, and A_0 is the cost incurred in rejecting a product. Thus, the objective with respect to the quadratic loss function is to achieve product response as close to the target as possible by choosing appropriate parameter settings and without modifying the product technology itself. On the other hand, the step loss function attributes no loss of quality to a product whose response lies within $\pm\delta$ of the target T, and hence no attempt is made to bring product response as close to the target as possible. Quality loss can be reduced in this approach only by improving technology – a more expensive alternative to parameter design.

Let $\mathbf{x} = [x_1, x_2, \ldots, x_K]$ be a vector denoting the values of K (controllable) factors which affect the product response y, and whose levels can be fixed by design. Since the

$$k_0 = \frac{A_0}{\delta^2}$$

$L(y)$ Quadratic loss function

A_0

Step function

$T - \delta$ $\quad T \quad$ $T + \delta$
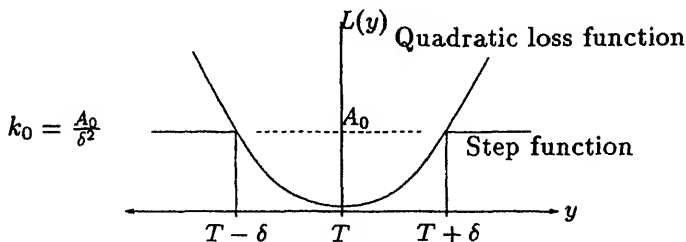
response $y(\mathbf{x})$ at the factor settings $\mathbf{x}$ varies randomly from sample to sample, resulting in a distribution of loss function values, an average (over the entire product population) of the loss function should be considered. For the quadratic loss function, the average loss function (assuming $k_0 = 1$) is the mean squared-error,

$$\text{MSE}(\mathbf{x}) = E[L(y(\mathbf{x}))] = \sigma^2(\mathbf{x}) + (\mu(\mathbf{x}) - T)^2,$$

where E is the expectation operator, and $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ are, respectively, the mean and variance of the response y parameterized by $\mathbf{x}$. For the mean squared-error type quality metric, quality is optimized by minimizing MSE($\mathbf{x}$) with respect to $\mathbf{x}$ taking values in a set Ξ (factor space).

Taguchi has proposed quality measures, called *signal-to-noise ratios*, for three different types of quality characteristics (e.g., Taguchi 1987; Phadke 1989). For a *smaller-the-better* type quality characteristic, where the target T to be achieved by the characteristic y is zero, it is defined as

$$SN_S(\mathbf{x}) = -10 \log_{10} E[y^2(\mathbf{x})].$$

For a *nominal-the-best* type characteristic, it is defined as

$$SN_T(\mathbf{x}) = 10 \log_{10}[\mu^2(\mathbf{x})/\sigma^2(\mathbf{x})],$$

and for the *larger-the-better* type characteristic, it is defined as

$$SN_L(\mathbf{x}) = -10 \log_{10} E[1/y^2(\mathbf{x})].$$

Taguchi's assertion is that maximization of these signal-to-noise ratios with respect to $\mathbf{x}$ would result in minimization of variability. For the *nominal-the-best* characteristics, he suggests identification of *signal factors* which affect the mean of y, but not its SN_T; the signal factors can be used to bring the mean to the target following the maximization of SN_T. Statisticians have devoted extensive effort to justify Taguchi's measures and his two-step approach, but have found only very special scenarios in which their use can be validated (León *et al* 1987; Box 1988). Where the quadratic loss function is appropriate, the use of the mean squared-error (MSE) would be ideal, provided it can be empirically modelled using experiments. The fact that signal-to-noise ratios have been reported to have worked in several case studies demonstrates that there is room for significant improvement in many currently used manufacturing processes and that even *ad hoc* methods can realize some of that improvement.

2.2 The scope for robust design

To demonstrate that variability can indeed be reduced by parameter design, consider the following illustrative example (taken from Boza *et al* 1994).

A circuit example: Consider the simple AC circuit shown in figure 2, where the power source $\mathcal{E}$ is known to operate at a tightly toleranced frequency f of either 50 or 60 Hz, and an rms value of 100 VAC with a tolerance of $\pm 10\%$. It is desired that the rms current I be as close to 10 amperes and with as little variability as possible. The design parameters are
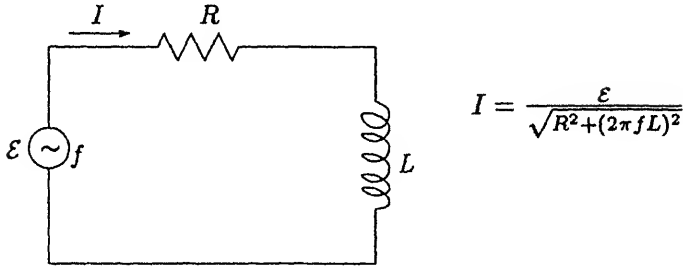
Figure 2. An AC circuit.

$$I = \frac{\mathcal{E}}{\sqrt{R^2 + (2\pi f L)^2}}$$

he nominal values of the resistance R and the inductance L, which are also toleranced at $\pm 10\%$ about their nominal values. It is assumed that $\mathcal{E}$, R and L are independent random ariables with means at their nominal values and the ($\pm 10\%$) tolerances at 3σ limits from ominal, e.g., the nominal value R_0 of R is such that

$$3\sigma_{R_0} = 0.1 R_0.$$

he frequency f is assumed to have zero variance about a nominal of either 50 or 0 Hz.

Using the Taylor series expansion about the nominal values of the circuit parameters, he current I can be expressed as

$$I = I_0 + \left.\frac{\partial I}{\partial \mathcal{E}}\right|_0 \Delta\mathcal{E} + \left.\frac{\partial I}{\partial R}\right|_0 \Delta R + \left.\frac{\partial I}{\partial L}\right|_0 \Delta L$$

$$+ \left.\frac{\partial I}{\partial f}\right|_0 \Delta f + \cdots \text{(higher-order terms)}, \tag{1}$$

vhere the subscript zero on I and the partial derivatives denotes their evaluation at the ominal parameter values $\mathcal{E}_0$, R_0, L_0 and f_0 ($\mathcal{E}_0 = 100$ VAC, $f_0 = 50$ or 60 Hz), and

$$\Delta\mathcal{E} = \mathcal{E} - \mathcal{E}_0, \quad \Delta R = R - R_0, \quad \Delta L = L - L_0, \quad \Delta f = f - f_0.$$

Jsing the assumptions that the parameters are independent and that the terms of third and igher orders are negligible, taking the expectation of I in (1) gives the mean value of the utput current as

$$\mu_I(\mathcal{E}_0, f_0, R_0, L_0) = I_0 + \frac{1}{2!}\left(\left.\frac{\partial^2 I}{\partial \mathcal{E}^2}\right|_0 \sigma_{\mathcal{E}_0}^2 + \left.\frac{\partial^2 I}{\partial R^2}\right|_0 \sigma_{R_0}^2\right.$$

$$\left. + \left.\frac{\partial^2 I}{\partial L^2}\right|_0 \sigma_{L_0}^2 + \left.\frac{\partial^2 I}{\partial f^2}\right|_0 \sigma_{f_0}^2\right) + \cdots$$

$$\simeq I_0 + \frac{1}{2}\left(\left.\frac{\partial^2 I}{\partial R^2}\right|_0 \sigma_{R_0}^2 + \left.\frac{\partial^2 I}{\partial L^2}\right|_0 \sigma_{L_0}^2\right). \tag{2}$$

obtaining (2), we have also used the facts that $(\partial^2 I/\partial \mathcal{E}^2) \equiv 0$ and $\sigma_f^2 \simeq 0$. For the minal values being considered, the bias $E(I) - I_0$ is negligible (less than 0.1%), and so

$$\sigma_I^2(\mathcal{E}_0, f_0, R_0, L_0) \simeq \left[\frac{\partial I}{\partial \mathcal{E}}\right]_0^2 \sigma_{\mathcal{E}_0}^2 + \left[\frac{\partial I}{\partial R}\right]_0^2 \sigma_{R_0}^2 + \left[\frac{\partial I}{\partial L}\right]_0^2 \sigma_{L_0}^2. \tag{3}$$

With the choice of $\mathcal{E}_0$ and f_0 not in the designer's control, the mean squared-error

$$\mathrm{MSE}_I(R_0, L_0) = \sigma_I^2(R_0, L_0) + (\mu_I(R_0, L_0) - T_I)^2,$$

(where $T_I = 10\mathrm{A}$ is the target for I) can be minimized with respect to the nominal values R_0 and L_0 of the series resistance and inductance to obtain a robust circuit design. Using the functional relationship between I and the circuit parameters given in figure 2, the following partial derivatives can be derived:

$$\frac{\partial I}{\partial \mathcal{E}} = \frac{I}{\mathcal{E}}; \qquad\qquad \frac{\partial^2 I}{\partial \mathcal{E}^2} = 0;$$

$$\frac{\partial I}{\partial R} = -\frac{I^3 R}{\mathcal{E}^2}; \qquad\qquad \frac{\partial^2 I}{\partial R^2} = \frac{I^3}{\mathcal{E}^2}\left(\frac{3R^2 I^2}{\mathcal{E}^2} - 1\right);$$

$$\frac{\partial I}{\partial L} = -(2\pi f)^2\frac{I^3 L}{\mathcal{E}^2}; \qquad \frac{\partial^2 I}{\partial L^2} = (2\pi f)^2\frac{I^3}{\mathcal{E}^2}\left(\frac{3(2\pi f)^2 L^2 I^2}{\mathcal{E}^2} - 1\right).$$

The contours of the mean squared-error versus the nominal values of R and L are shown in figure 3 for both nominal frequencies. It can be seen that the surface has ridge-like behaviour, and, hence, for large ranges of R and L values, the mean squared-error is very close to its minimum.

Boza *et al* (1994) assumed an initial design setting of $(R_0, L_0) = (5\Omega, 0.02\mathrm{H})$, and obtained a final design of $(R_0, L_0) = (8\Omega, 0.017\mathrm{H})$. Their objective, however, was only the minimization of the variability σ_I^2 which was a sufficient goal as the customer-specified tolerance of 2.5 amperes on the targeted I was not violated. However, to achieve good robustness at both nominal frequencies, figure 4 suggests a design near $(R_0, L_0) = (9.43\Omega, 0.01\mathrm{H})$, when at both frequencies an MSE = $0.21\mathrm{A}^2$ is achieved.

Table 1 gives a summary of the circuit's performance at the initial design, the design obtained by Boza *et al* (1994), the designs at which MSE is optimum at either frequency,
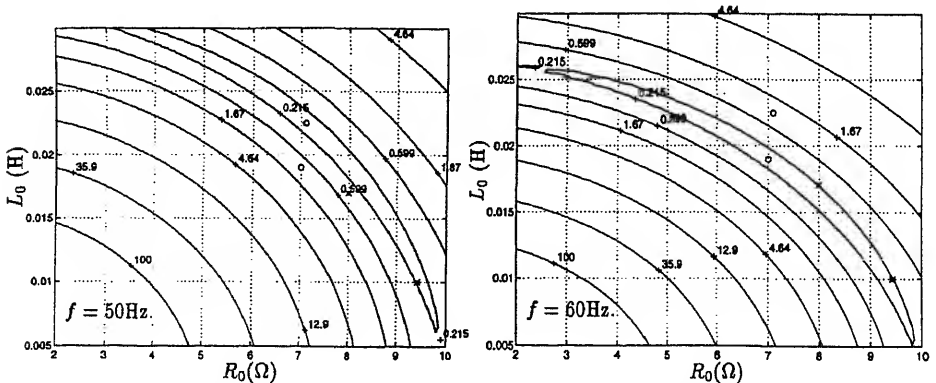
Figure 3. Contour-plots of MSE_I versus R_0 and L_0 at 50 and 60 Hz. The '$*$' marks our final design, the '$\times$' marks the design obtained by Boza *et al* (1994), and the 'o's mark the
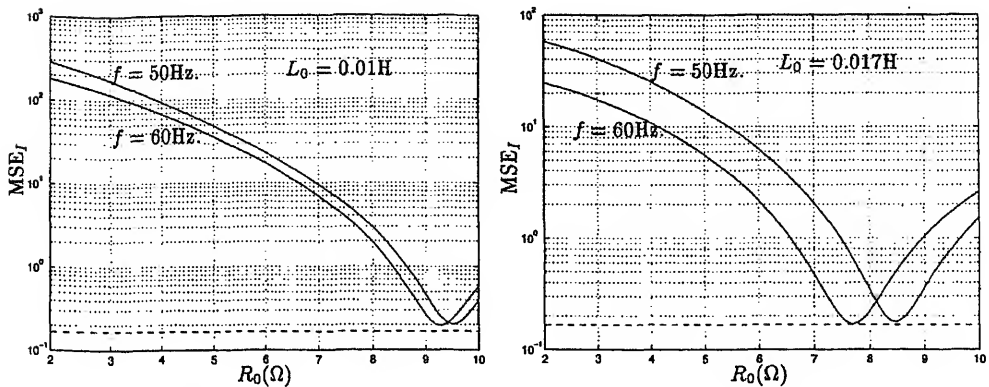
Figure 4. MSE profiles versus R_0 at two fixed values of L_0; the dashed line indicates the minimum achievable MSE under either frequency scenario.

and at our suggested design. Table 1 also gives the values of the *capability index* C_{pk} at either frequency for each parameter design. C_{pk} is defined as (e.g., Montgomery 1991a),

$$C_{pk} = \min \left\{ \frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma} \right\},$$

where USL and LSL are respectively the upper and lower specification limits for the response (for this example, $USL_I = 12.5A$ and $LSL_I = 7.5A$). If the above expression returns a negative number, C_{pk} is taken to be zero. For Six-Sigma Quality (pioneered by Motorola Inc; Harry 1988), a C_{pk} value should be greater than or equal to 1.5. For this example, robust design has indeed achieved this goal. Further performance gains are possible only by technological improvement, e.g., by improving the component tolerances.

The above example demonstrates that in order for the product to function equally well under two frequency scenarios, a *compromise* design should be used. While this compromise design was relatively easy to obtain using the contour plots of figure 3 and the MSE profiles in figure 4, a visual approach is not possible if the number of design factors or the number of objectives exceeds two. The problem of multiple objectives is examined in the next section.

2.3 *Experimental response modelling*

In the example considered in the previous sub-section, a physical relationship between the response of interest and the design factors was available. In a general manufacturing

Table 1. Performance of AC circuit for different parameter designs.

R_0 (Ω)	L_0(H)	$f = 50$ Hz				$f = 60$ Hz			
		μ_I (A)	σ_I (A)	MSE_I (A^2)	C_{pk}	μ_I (A)	σ_I (A)	MSE_I (A^2)	C_{pk}
5	0.02	12.458	0.513	6.302	0.03	11.058	0.462	1.333	1.04
7.1	0.0225	9.984	0.408	0.166	2.03	9.06	0.371	1.053	1.39
7	0.019	10.874	0.446	0.962	1.22	9.988	0.408	0.166	2.03
8	0.017	10.4	0.435	0.349	1.61	9.759	0.401	0.219	1.88
9.43	0.01	10.069	0.452	0.210	1.79	9.854	0.436	0.211	1.80

process, this would rarely be the case. Consequently, empirical models for the metrics are obtained. An experiment consists of a systematic variation of the factors and recording of the response(s) defining the quality metric(s) to enable reliable estimation of the models. For an experiment of N runs, the procedure involves obtaining (noisy) measurements $y(\mathbf{x})$ of a response at N distinct values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ in the space Ξ of the design factors. In Taguchi experiments, since the objective of interest is an aggregate measure, such as the signal-to-noise ratio of a product/process characteristic, which summarizes the variability of quality characteristic for a given combination of design-factor settings, an experiment involves variation of not only the *design factors*, but also a systematic variation of *noise factors* that contribute to the variability in the product's response. Noise factors are variables that are uncontrollable in the real environment but can be manipulated in the experimental set-up. Thus, a Taguchi experiment consists of (i) an *inner array* which lists the design factor-level combinations, and (ii) an *outer array* that lists the noise factor-level combinations for each design factor-level combination. If m is the number of combinations of noise factors chosen for an outer array, an experiment yields m replicate measurements for each of the N runs. What the best placement of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ should be depends on the assumed model-type for $\eta(\mathbf{x})$ (e.g., linear, quadratic etc.), and on practical considerations. The problem of design of experiments has been extensively researched during this century by applied statisticians, and there exists a rich body of literature documenting numerous design types and approaches (see, for example, Box *et al* 1978; Montgomery 1991b, and the bibliographies therein). In this article, we shall sidestep this issue by assuming that, for a given problem, an appropriate design has been identified. However, for the purposes of this paper, we mention one useful and frequently used design-type: the *fractional-factorial design*. Most of the *orthogonal-array* designs compiled and recommended by Taguchi are of the fractional-factorial type. For example, consider an experiment involving the study of the effects of three factors x_1, x_2 and x_3 on the response y. If each factor is constrained to take only two fixed levels, denoted by 1 and 2, during the experiment, the maximum number of distinct runs N is 8, so that all combinations of the levels of each factor are tested. This arrangement, shown in the array L_8 in figure 5, is a two-level *full-factorial* design. If only four of the eight possible combinations are run,

L_8

Run	Factor			
no.	x_1	x_2	x_3	y
1	1	1	1	y_1
2	1	1	2	y_2
3	1	2	1	y_3
4	1	2	2	y_4
5	2	1	1	y_5
6	2	1	2	y_6
7	2	2	1	y_7
8	2	2	2	y_8

L_4

Run	Factor			
no.	x_1	x_2	x_3	y
1	1	1	1	y_1
2	1	2	2	y_2
3	2	2	1	y_3
4	2	1	2	y_4

as in the array L_4 in figure 5, the resulting design is of the fractional-factorial type. A useful feature of fractional-factorial designs is their orthogonality, i.e., the analysis of data with respect to each factor can be done using only the column(s) corresponding to that factor. Orthogonal arrays (which include some fractional-factorial designs) also possess the following key features:

- each level of a factor appears in an equal number of runs (factor-level combinations); this is called the balancing property – it ensures unbiased estimation of all column effects;
- the effects measured by any two columns are mutually orthogonal (uncorrelated);
- the effect measured by a column may be aliased (confounded) with the combined effect (interaction) of two or more columns.

Orthogonal arrays are normally used in Taguchi's off-line design procedure for the *inner array* which determines the combinations of the design factor-levels. In addition, their use has also been recommended in the *outer array* which determines the combinations of noise factor-levels, even though the statistical implications of doing are debatable (Montgomery 1991a).

2.4 *Data analysis for orthogonal-array experiments*

In this sub-section, we briefly review the common data analysis techniques for experimental data: analysis of means, analysis of variances (ANOVA), and multiple linear regression. These techniques are used to estimate the functional relationship between the factors and a response, and to assess the statistical significance of the estimated relationship. While ANOVA and regression are the common tools of choice, the analysis of means is often convenient for measuring factor effects when an orthogonal-array experiment is used. The following results are given with respect to analysis of experimental data using an orthogonal array.

If a trial (or run) corresponds to the level combination $(x_1, x_2, \ldots, x_K)$ of the K factors (denoted by $\mathbf{x}$), then we shall assume that a response y can be expressed as

$$y(\mathbf{x}) = \eta(\mathbf{x}) + \varepsilon,$$

where $\eta(\mathbf{x})$ is an unknown deterministic function — the mean response corresponding to this factor-level combination, and ε is a zero-mean random noise with variance σ^2. Here, the response y is used to denote measurements of either the quality characteristic or any suitable loss function derived from it, e.g., the MSE or a signal-to-noise ratio. In either case, we assume that ε is independent with constant variance over all experimental runs. The task of experimental modelling is eventually to estimate the functional relationship $\eta(\mathbf{x})$.

2.4a *Analysis of means and variances* : Consider a balanced experiment in which each factor-level combination appears in an equal number of trials, and, consequently, each of the L_k levels of factor x_k appears in an equal number, n_k, of trials, i.e., $n_k L_k = N$. The number of trials N need not be equal to the total number of possible factor-level combinations, but in a factorial design, for example, N must divide $\prod_{k=1}^{K} L_k$. We assume

that we take m replicates per trial and, hence, have Nm observations of the response variable: y_{ij}, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, m$ (note that if y denotes an aggregate or summary measure such as an MSE or a signal-to-noise ratio, we would have $m = 1$ since replicated measurements of the quality characteristic would already have been used to compute y). Now, the following sample averages and sums of squares may be defined:

$$\bar{y} = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij}, \tag{4}$$

$$\bar{y}_i = \frac{1}{m} \sum_{j=1}^{m} y_{ij}, \quad i = 1, \ldots, N, \tag{5}$$

$$\bar{y}_{k,l} = \frac{1}{n_k m} \sum_{\{i: l_k = l\}} \sum_{j=1}^{m} y_{ij}, \quad k = 1, 2, \ldots, K; \quad l = 1, 2, \ldots, L_k, \tag{6}$$

$$S = \sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \bar{y})^2, \tag{7}$$

$$S_k = m n_k \sum_{l=1}^{L_k} (\bar{y}_{k,l} - \bar{y})^2, \quad k = 1, 2, \ldots, K, \tag{8}$$

$$S_e = \sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2, \tag{9}$$

$$S_r = S - \sum_{k=1}^{K} S_k = S_e + S_{LOF}, \tag{10}$$

where $\bar{y}$ is the overall mean, $\bar{y}_i$ is the average response in the ith trial, $\bar{y}_{k,l}$ is the average response due to factor k taking level l, $\{i : l_k = l\}$ is the set of indices denoting trial numbers in which factor k takes level l. S is the overall sum of squares (corrected for the mean), S_k is the sum of squares contributed by the variation due to changes in the levels of factor k, and S_e is the sum of squares due to pure random error (it accounts for the within-run variations). S_r, known as the residual sum of squares, accounts for the contribution S_{LOF} to the variability in y due to lack of fit (unaccounted-for effects, e.g., interactions), and due to pure random error (S_e). The first three of the above set of equations constitute the analysis of means (ANOM) which enables measurement of effects of factors. The remaining four equations are used for ANOVA, which enables the identification of significant factor effects that should be included in the model for η. In (6), the mean responses have been computed only for each of the factors taking a given level (main effects) assuming a linear (additive) model for η. But a factorial design and some orthogonal arrays may permit estimation of not just the main effects but also interaction effects that measure the combined effect of two or more factors on the response. A two-level factorial design, for example, can measure as many effects (main or interaction) as the number of columns in the array. The computation of these effects is similar to that given in (6) since exclusive (orthogonal) columns in the array can be assigned to measure these effects provided N is large enough to accommodate their estimation (see Box $et\ al$ 1978; Montgomery 1991b). In general, if p (main and interaction) effects can be measured by an orthogonal array ($p \leq N - 1$), a

prediction equation for η can be obtained as,

$$\hat{\eta}(l_1, l_2, \ldots, l_K) = \bar{y} + \sum_{i=1}^{p} (\bar{y}_{i,l_i} - \bar{y}), \tag{11}$$

where $\hat{\eta}(l_1, l_2, \ldots, l_K)$ is the estimated response at the factor-level combination $(l_1, l_2, \ldots, l_K)$ of the K factors; the levels of the interaction effects are determined from those of the corresponding interacting factors. The *goodness of fit* of the model can be measured by the R^2 value defined as

$$R^2 = \left(\sum_{i=1}^{p} S_i \right) \Big/ S.$$

The maximum possible R^2 is $(S - S_e)/S$; if the number of replications $m = 1$, the maximum possible R^2 is 1.

2.4b *Multiple linear regression*: An ANOVA is usually accompanied by least-squares regression to estimate the response-surface model for η. ANOVA gives the significant effects that can be included in a polynomial model for η. That is, $\eta(\mathbf{x})$ is assumed to have the functional form

$$\eta(\mathbf{x}) = \boldsymbol{\theta}' \mathbf{z}(\mathbf{x}),$$

where

$$\mathbf{z}(\mathbf{x}) = [1 \; x_1 \; \ldots \; x_k \; x_1^2 \; \ldots \; x_k^2 \; x_1 x_2 \; \ldots]'$$

is a $p \times 1$ vector containing 1 as the first element, and powers and cross-products of $x_1, x_2, \ldots, x_k$, that are found significant from ANOVA, as the remaining $p - 1$ elements. Assuming $m = 1$, if $\mathbf{z}(\mathbf{x}_i)'$, $i = 1, \ldots, N$, corresponding to the N runs, constitute the rows of the $N \times p$ matrix X, and if $\mathbf{y} = [y_1, \ldots, y_N]'$ denotes the vector of N independent measurements of the response, least-squares regression gives the estimates (Anderson 1984),

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1} X' \mathbf{y}, \tag{12}$$

$$\hat{\eta}(\mathbf{x}) = \hat{\boldsymbol{\theta}}' \mathbf{z}(\mathbf{x}), \tag{13}$$

for the model coefficients and the response surface.

2.4c *M responses*: Multiple linear regression can be extended in a straightforward manner to the case of simultaneous modelling of M responses using design matrix X. Now M vectors of coefficients, represented as the M columns of the $p \times M$ matrix Θ in the model,

$$\mathbf{y}(\mathbf{x}) = \Theta' \mathbf{z}(\mathbf{x}) + \boldsymbol{\varepsilon},$$

where $\Theta = [\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_M]$ are to be estimated, the measurements $\mathbf{y}$ and the noise $\boldsymbol{\varepsilon}$ are M-vectors, and $\boldsymbol{\varepsilon}$ is assumed to be zero-mean with $M \times M$ covariance matrix Σ. Now, based on the measurement data

$$Y = X\Theta + \mathbf{E},$$

where Y and $\mathbf{E}$ are $N \times M$, and X is $N \times p$, the least-squares estimates are

$$\hat{\Theta} = (X'X)^{-1}X'Y, \tag{14}$$
$$\hat{\eta}(\mathbf{x}) = \hat{\Theta}'\mathbf{z}(\mathbf{x}). \tag{15}$$

The least-squares estimates satisfy the following statistics (Anderson 1984)

$$E(\hat{\theta}_j) = \theta_j \quad \text{(unbiased)}$$
$$\text{Cov}(\hat{\theta}_i, \hat{\theta}_j) = \sigma_{ij}(X'X)^{-1}$$
$$E(\hat{\eta}(\mathbf{x})) = \Theta'\mathbf{z}(\mathbf{x}) = \eta(\mathbf{x}) \quad \text{(unbiased)} \tag{16}$$
$$\text{Cov}(\hat{\eta}(\mathbf{x})) = \mathbf{z}(\mathbf{x})'(X'X)^{-1}\mathbf{z}(\mathbf{x})\Sigma. \tag{17}$$

σ_{ij} is the ijth element of Σ ($\sigma_{jj} = \sigma_j^2$). Also,

$$\hat{\Sigma} = [1/(N - p)]Y'[I_N - X(X'X)^{-1}X']Y, \tag{18}$$

where I_N is the $N \times N$ identity matrix, and $\hat{\Sigma}$ is the unbiased estimator of Σ provided Y is of rank M ($M \le N - p$).

Under normality assumptions, the estimates are distributed according to the following two independent (multivariate) distributions

$$(\hat{\theta}_1' \, \hat{\theta}_2' \cdots \hat{\theta}_M')' \sim \mathcal{N}((\theta_1'\theta_2' \cdots \theta_M')', \Sigma \cdot (X'X)^{-1}), \tag{19}$$
$$(N - p)\hat{\Sigma} \sim W(\Sigma, N - p), \tag{20}$$

where $\Sigma \cdot (X'X)^{-1}$ is a scalar product of the two matrices (an $mp \times mp$ matrix with each element of Σ multiplied by the matrix $(X'X)^{-1}$). In (20), $W(\Sigma, N - p)$ is a Wishart distribution with $(N - p)$ degrees of freedom, parameterized by the matrix Σ (see Anderson 1984). The diagonal elements of a $W(\Sigma, N - p)$-distributed matrix are chi-squared distributed random variables with $N - p$ degrees of freedom.

In the rest of this article, the use of prediction models, obtained either from least-squares regression or analysis of means, for optimizing quality criteria will be considered.

3. Multiobjective optimization

Consider now the robust design of a product/process with respect to M quality metrics. Ignoring for now their estimation using experimentation, suppose the response functions $\eta(\mathbf{x}) = [\eta_1(\mathbf{x}) \, \eta_2(\mathbf{x}) \cdots \eta_M(\mathbf{x})]'$ are perfectly known. The design problem then involves obtaining the process variable settings $\mathbf{x}^* \in \Xi$ that simultaneously result in the most desirable compromise of the m responses. As against single-objective optimization, in which the optimum response is a unique value (maximum or minimum response), 'optimization' with respect to multiple objectives refers to the attainment of any one solution in a set termed the *nondominated solution set*. A solution is said to be nondominated (or Pareto-optimal, or noninferior) if no other solution is at least as good as this solution with respect to every objective and better than this solution with respect to at least one
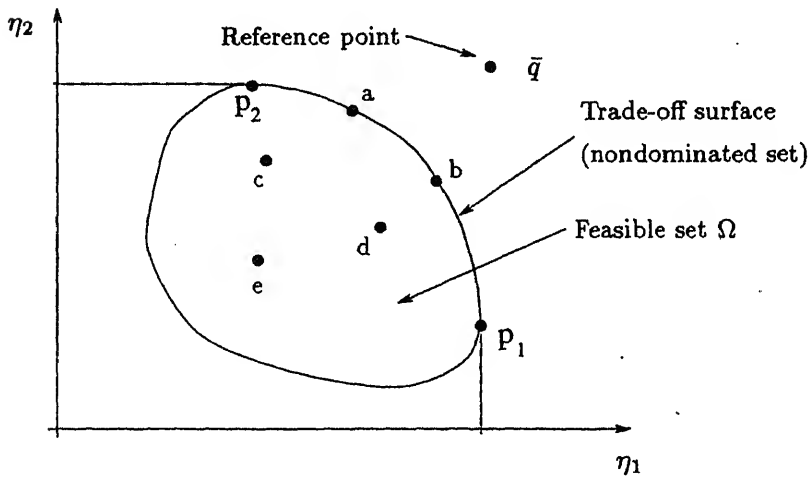
Figure 6. The concept of nondominated solutions (in a two-dimensional objective space).

objective (see Chankong & Haimes 1983). One can use the partial ordering available in the M-dimensional objective space $\mathcal{R}^M$ to identify the nondominated set by eliminating each of those feasible points that is definitely worse than at least one other feasible point (with respect to all M objectives). The concept of nondominated points is illustrated in figure 6 for the case of maximization of two objectives ($M = 2$), where the set of attainable (feasible) objective-pairs is shown as a closed set Ω. Seven points in this set: **a**, **b**, **c**, **d**, **e**, $\mathbf{p}_1$ and $\mathbf{p}_2$, are also marked for the purpose of illustration. Point **a** dominates **c** and **e** since **a** is better than **c** as well as **e** with respect to both objectives, η_1 and η_2. Similarly, **b** dominates **d** and **e**. Points **a**, **b**, $\mathbf{p}_1$, $\mathbf{p}_2$, and all other points on that segment of the boundary of the feasible region Ω between $\mathbf{p}_1$ and $\mathbf{p}_2$, are nondominated, while **c**, **d**, **e**, and all other points inside the feasible region are dominated.

The concept of domination or partial ordering in an M-dimensional objectives space $\mathcal{R}^M$ can be mathematically defined using the concept of a *positive cone D* (any closed, convex, proper cone):

$$\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{R}^M, \mathbf{q}_1 \leq \mathbf{q}_2 \iff \mathbf{q}_2 - \mathbf{q}_1 \in D.$$

In figure 7, since the point $\mathbf{q}_2 - \mathbf{q}_1$ is in D, $\mathbf{q}_2$ dominates $\mathbf{q}_1$, and since both individually lie in D as well, both dominate the zero-vector $\mathbf{0}$. A nondominated objective (or *D-maximal* or Pareto-optimal objective) $\check{\mathbf{q}}$ in Ω is defined by

$$\check{\mathbf{q}} \in \Omega \text{ is } D\text{-maximal} \iff \Omega \cap (\check{\mathbf{q}} + \tilde{D}) = \emptyset.$$

The selection of any particular nondominated solution from the set of all nondominated solutions in Ω must be qualified by the preferences of the decision-maker (process- or quality-engineer), for the choice of any one nondominated point over the others implies a *trade-off* of one or more objectives for a gain in another objective. There exist several different, but closely related, methods of incorporating a decision-maker's preferences to search for the final solution (see, for example, Zeleny 1982, Steuer 1989). In one such method, called the *method of reference points* (Wierzbicki 1980), a reference point (a vector of desired objectives) is specified by the decision maker. The method maximizes
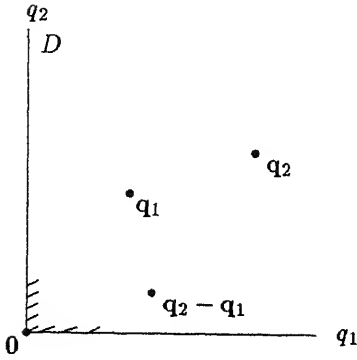
Figure 7. Positive cone D in a two-dimensional ($M = 2$) objectives space $\mathcal{R}^2$.

a *scalarizing function* called the *achievement function* that guarantees a nondominated solution. This is introduced in the next section.

4. Multiobjective optimization of response surfaces

Most designed experiments enable estimation of models for the quality metrics over a continuous region Ξ using least-squares regression. In this case, response surface models (Box & Draper 1987) are available for optimization, and nonlinear programming methods can be used. For multiobjective optimization, the *reference point method* can be applied by maximizing a scalarizing function (called the *achievement function*) of the M response models over the feasible space. One possible scalarizing function is (Lewandowski *et al* 1989)

$$s(\mathbf{q}, \overline{\mathbf{q}}) = \min_{i \in \{1,2,\ldots,M\}} \alpha_i (q_i - \overline{q}_i) + \beta \sum_{j=1}^{M} \alpha_j (q_j - \overline{q}_j), \tag{21}$$

defined for a general point $\mathbf{q}$ in the objectives space $\Omega \subset \mathcal{R}^M$, and for a reference-point $\overline{\mathbf{q}}$. β and α_i, $i = 1, \ldots, M$ are positive constants, whose values are to be fixed by the decision-maker.

Figure 8 illustrates the concept of the achievement function in (21) in a two-objectives space, where the two objectives q_1 and q_2 are to be maximized subject to the constraint that they lie in the set Ω of attainable objective-pairs. The reference point $\overline{\mathbf{q}}$ that is chosen in this figure is unattainable; in general, it could lie inside Ω and the reference-point method would still return a nondominated solution. The 'ideal' point (Khuri & Conlon 1981), whose coordinates are the maxima of the individual objectives, is also shown, denoted here by $\mathbf{q}_I$. (In an example that we have considered later in this section, the ideal point is chosen as the reference point.) The contours of the achievement function in (21) are shown for two values: $s = 0$, and $s = s_2$ ($s_2 < 0$). The shape of these contours is also the shape of the chosen domination cone D_β, e.g., all points right of the contour corresponding to $s = 0$ dominate the reference point $\overline{\mathbf{q}}$. The boundary $\overline{\Omega}$ of the attainable set Ω indicated by the thick line comprises the nondominated set, i.e., no other point in Ω dominates a point in this set with respect to both objectives. The point $\mathbf{q}^*$ is the multiobjective optimum solution obtained by the maximization of the chosen achievement function over $\mathbf{q} \in \Omega$ (the
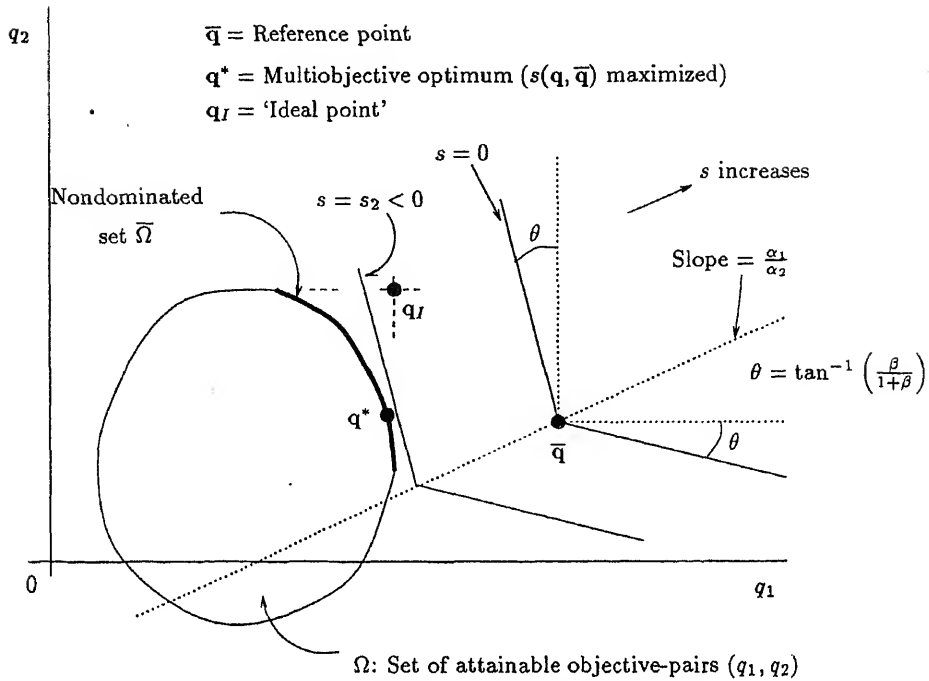
Figure 8. Concepts of reference point and achievement function for a two-objective maximization case.

s-contour is tangent to the set Ω at $\mathbf{q}^*$). β is related to the angle θ as shown in figure 8. It determines the order that a decision-maker wishes to assign to points that do not dominate one another. The parameters α_1 and α_2 determine the slope of the dividing plane between the two piecewise-linear regions for s.

The optimization problem can now be expressed as

$$\max_{\mathbf{q}\in\Omega} s(\mathbf{q}, \overline{\mathbf{q}}) \tag{22}$$

or, using $\mathbf{q} = \eta(\mathbf{x})$,

$$\max_{\mathbf{x}\in\Xi} s(\eta(\mathbf{x}), \overline{\mathbf{q}}). \tag{23}$$

When a unique, deterministic mapping $\eta : \Xi \longrightarrow \mathcal{R}^M$ is not known, and response surface models have to be used, one may instead maximize

$$s(\hat{\eta}(\mathbf{x}), \overline{\mathbf{q}}) = \min_{i\in\{1,2,\dots,M\}} \alpha_i (\hat{\eta}_i(\mathbf{x}) - \overline{q}_i) + \beta \sum_{j=1}^{M} \alpha_j (\hat{\eta}_j(\mathbf{x}) - \overline{q}_j). \tag{24}$$

The issue that the objectives are no longer deterministic, but involve modelling uncertainty, can be handled as follows (see Mathur & Pattipati 1995).

The maximum position $\mathbf{x}^*$ of $s(\hat{\eta}(\mathbf{x}), \overline{\mathbf{q}})$ is declared as the 'multiobjective optimum' but, in addition, a region similar to a confidence region is associated with it. This region is obtained via Monte Carlo simulations about the estimated models using the least-squares estimates and their distributions. Since, from (19),

$$(\theta_1' \, \theta_2' \cdots \theta_M')' - (\hat{\theta}_1' \, \hat{\theta}_2' \cdots \hat{\theta}_M')' \sim \mathcal{N}(0, \, \Sigma \cdot (X'X)^{-1}),$$

the model coefficients $\theta_1, \theta_2, \ldots, \theta_M$ may be generated using the normal distribution,

$$\mathcal{N}((\hat{\theta}_1' \hat{\theta}_2' \cdots \hat{\theta}_M')', \hat{\Sigma} \cdot (X'X)^{-1}), \tag{25}$$

which is parameterized by $\hat{\Sigma}$ instead of the (unknown) Σ. Also, since $(N-p)\hat{\Sigma}$ is itself distributed as a Wishart distribution $W(\Sigma, N-p)$ (see, for example, Anderson 1984), one might generate matrices using $W(\hat{\Sigma}, N-p)$ and use these to parameterize the above normal distribution. The multiobjective optima of the models so generated can then be obtained for all the runs and plotted on a scatter plot to visualize the region of distribution of the optima. For cases involving more than two or three input variables, where visualization of such a region is not easy, a Monte Carlo significance test procedure (Barnard 1963) may be used to compute a boundary for this region.

Monte Carlo significance test: Let $\mathbf{x}_i, i = 1, 2, \ldots, N_s$, be the positions of the optima obtained from N_s independent Monte Carlo simulations. Let these points be realizations of independent random vectors with distribution $p(\mathbf{x})$ which is unknown. Let $t(\mathbf{x})$ be a test criterion for the hypothesis that the true optimum is distributed according to $p(\mathbf{x})$. We denote these statistics by $t_1, t_2, \ldots, t_{N_s}$ for the N_s runs. Now, for a level of significance α $(0 < \alpha < 1)$, the region in which the true optimum lies can be obtained as

$$\{\mathbf{x} : t(\mathbf{x}) < t_{(\lfloor N_s(1-\alpha)\rfloor)}\}, \tag{26}$$

where $t_{(j)}$ is the statistic of rank j obtained after sorting $\{t_i\}$ in ascending order, and $\lfloor N_s(1-\alpha)\rfloor$ is the largest integer less than or equal to $N_s(1-\alpha)$. We shall refer to the region defined by the set in (26) as the 'significance region.'

The choice of a statistic t depends on the distribution $p(.)$ of the $\mathbf{x}_i$, which is unknown. The ideal (but intractable) approach would be to obtain a nonparametric estimate of the distribution function based on the N_s measurements $\mathbf{x}_i$. However, for some applications, it is possible to obtain reasonably good approximate regions based on a statistic that utilizes the sample estimates of the first few higher-order statistics: mean, covariance, (multivariate) skewness, kurtosis, and so on. Let $\bar{\mathbf{x}}$ be the sample mean, and S be the sample covariance matrix. In addition, defining $\mathbf{u}_i = S^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iK})'$, let $r_{lmn}, l, m, n = 1, 2, \ldots, K$, be the lmnth element of the sample third-order cumulant (skewness) tensor:

$$r_{lmn} = \frac{1}{N_s} \sum_{i=1}^{N_s} u_{il} u_{im} u_{in},$$

and let $\kappa_{lmnp}, l, m, n, p = 1, 2, \ldots, K$, be the $lmnp$th element of the sample fourth-order cumulant (kurtosis) tensor:

$$\frac{1}{N_s} \sum_{i=1}^{N_s}$$

Table 2. Experimental design and measured response data (example).

Run	Design (original) x_1 (mM)	x_2 (mM)	Design (coded) x_1	x_2	Responses $y_1 (kg)$	y_2	y_3	y_4 (mm)
1	8.0	6.5	-1	-1	2.48	0.55	1.95	0.22
2	34.0	6.5	1	-1	0.91	0.52	1.37	0.67
3	8.0	25.9	-1	1	0.71	0.67	1.74	0.57
4	34.0	25.9	1	1	0.41	0.36	1.20	0.69
5	2.6	16.2	-1.414	0	2.28	0.59	1.75	0.33
6	39.4	16.2	1.414	0	0.35	0.31	1.13	0.67
7	21.0	2.5	0	-1.414	2.14	0.54	1.68	0.42
8	21.0	29.9	0	1.414	0.78	0.51	1.51	0.57
9	21.0	16.2	0	0	1.50	0.66	1.80	0.44
10	21.0	16.2	0	0	1.66	0.66	1.79	0.50
11	21.0	16.2	0	0	1.48	0.66	1.79	0.50
12	21.0	16.2	0	0	1.41	0.66	1.77	0.43
13	21.0	16.2	0	0	1.58	0.66	1.73	0.47

mM = millimolar

where $\delta_{ij} = 1$, if $i = j$; otherwise, it is zero. Higher-order cumulants can similarly be computed. If the skewness, kurtosis, and higher-order cumulants are negligible, it may suffice to use the statistic

$$t_1(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}})' S^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{u}'\mathbf{u}. \tag{27}$$

In general, a statistic based on the likelihood function involving the higher-order terms in an Edgeworth expansion (see, for example, Kolassa 1994) could be used:

$$t_2(\mathbf{x}) = -\phi(\mathbf{u}) \left(1 + \frac{1}{6} \sum_{l,m,n=1}^{K} r_{lmn} H_{lmn}(\mathbf{u}) \right.$$
$$\left. + \frac{1}{24} \sum_{l,m,n,p=1}^{K} \kappa_{lmnp} H_{lmnp}(\mathbf{u}) + \cdots \right), \tag{28}$$

where $\phi(\mathbf{u})$ is the K-variate standard Gaussian density function (for mean $\mathbf{0}$ and covariance matrix I). $H_{lmn}(\mathbf{u})$ and $H_{lmnp}(\mathbf{u})$ are, respectively, the third-order and fourth-order K-dimensional Hermite polynomials defined as

$$H_{lmn}(\mathbf{u}) = \frac{(-1)^3}{\phi(\mathbf{u})} \frac{\partial^3 \phi(\mathbf{u})}{\partial u_l \partial u_m \partial u_n}$$
$$= u_l u_m u_n - (\delta_{lm} u_n + \delta_{mn} u_l + \delta_{nl} u_m),$$

and

$$H_{lmnp}(\mathbf{u}) = \frac{(-1)^4}{\phi(\mathbf{u})} \frac{\partial^4 \phi(\mathbf{u})}{\partial u_l \partial u_m \partial u_n \partial u_p}$$
$$= u_l u_m u_n u_p - (\delta_{np} u_l u_m + \cdots + \delta_{mp} u_l u_n)$$
$$+ (\delta_{lm} \delta_{np} + \delta_{ln} \delta_{mp} + \delta_{lp} \delta_{mn}).$$
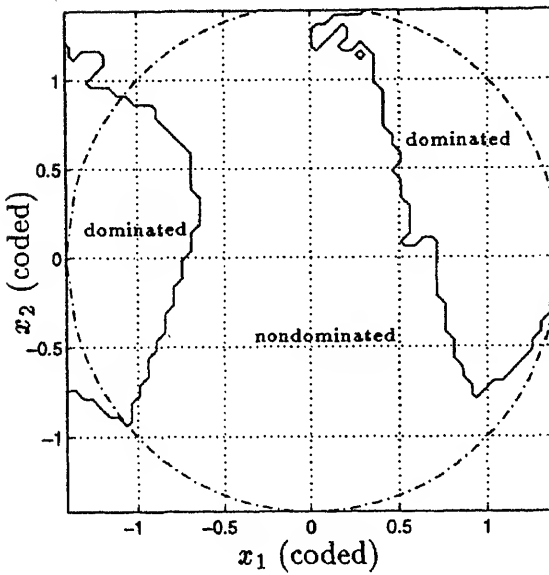
Figure 9. Regions corresponding to nondominated solutions and dominated solutions with respect to all four responses based on the estimated models (example).

The higher-order polynomials have similar extensions.

To summarize the above, the procedure involves the following steps:

(1) Using the experimental data for the responses, obtain regression models and estimates of the covariance matrices of the model coefficients using (15) and (18);

(2) Compute the achievement function (24) over the region Ξ using the predicted responses and the reference point input by the user; obtain the multiobjective optimum settings by maximizing the achievement function over the region Ξ;

(3) Obtain several independent sets of models by Monte Carlo simulations using the least-squares estimates of the statistics in the generating distribution (25); for each realization, repeat step 2 to obtain a sample of points $\{x_i\}$;

(4) Apply the statistic in (28) (or (27) if adequate) to the sample optima $\{x_i\}$ of step 3 to obtain a region of the form (26).

The feasible region Ξ in the factor space is recommended to be (two or three times) larger than the experimental design region in order to ensure that the sample points obtained

Table 3.   The second degree regression models (example).

Model term	Regression coefficients			
	y_1	y_2	y_3	y_4
int.	1.526	0.660	1.776	0.468
x_1	−0.575	−0.092	−0.250	0.131
x_2	−0.524	−0.010	−0.078	0.073
x_1^2	−0.171	−0.096	−0.156	0.026
x_2^2	−0.098	−0.058	−0.079	0.024
$x_1 x_2$	0.318	−0.070	0.010	−0.083
R^2	0.95	0.98	0.98	0.95

(a) (b)

Figure 10. Distribution of the dual-objective optimum (for y_2 and y_3) for the example; the '+'s show the positions of the optima of the individual (estimated) responses. The region of significance is obtained using the statistic t_1 in (**a**) and t_2 (with the first four orders of moments) in (**b**).

via simulations lie inside Ξ (see Mathur & Pattipati 1995); for if a large number of such points lie on the boundary of Ξ, the estimation of the moments (and cumulants) used in the statistic t could incur large errors. In the final analysis, all inferences can be confined to the experimental region. The region of significance should be interpreted as follows: if it includes the design centre (current settings), there is not enough evidence to suggest a change in the factor settings; if it does not, but is large compared to the size of the experimental region, or if it lies outside the experimental region, further experimentation is needed.

4.1 *Example*

The following example was originally discussed by Khuri & Conlon (1981) (also see Mathur & Pattipati 1996). In this case study of a dialyzed whey-protein-concentrate (WPC) gel system, the effects of two inputs: concentrations of cysteine (x_1) and calcium chloride (x_2), on four responses measuring the textural and water-holding characteristics of the WPC gel were studied. The four texture characteristics are: hardness (y_1), cohesiveness (y_2), springiness (y_3), and compressible water (y_4). The goal of the problem is to maximize the measures of all these characteristics. A 13-point rotatable central-composite-design experiment (Montgomery 1991b) with five centre points (for uniform precision) was conducted to vary the input variables. The experiment design and the measured values of the two responses are recorded in table 2, reproduced from Khuri & Conlon (1981).

(a)

(b)
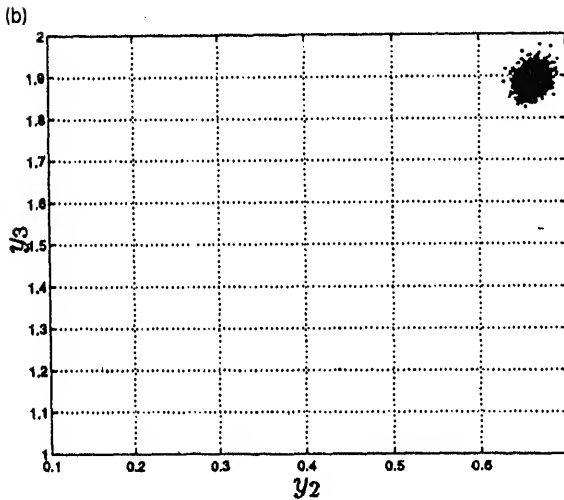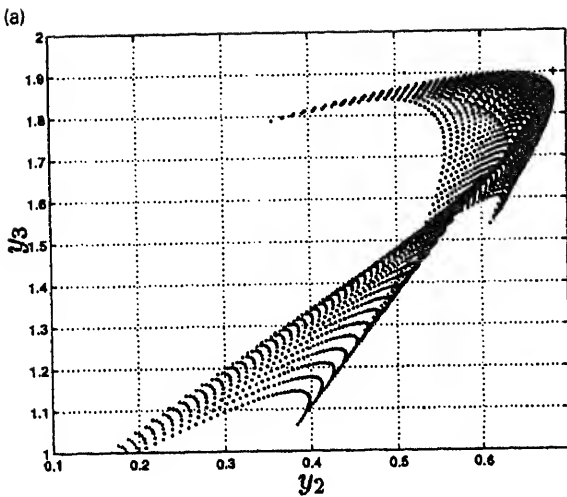
Figure 11. (Example, two-response case) (a). Attainable objective pairs (of y_2 and y_3) based on estimated models. The '+' marks the chosen reference point. (b). Scatter of the dual-objective optima obtained from 1000 Monte Carlo simulations of the models.

Only y_2 and y_3 considered: Considering all four responses, it turns out that the current operating point $\mathbf{x} = (0, 0)$ is efficient (nondominated). That is, no further improvement in any response is possible without trade-off in at least one other response. In fact, a large area of the region $x_1^2 + x_2^2 \leq 2$ maps to the nondominated set in the objective space (see figure 9; this efficient set is based on the predicted responses). So, for illustrative purposes, we first consider only two of the four responses: y_2 and y_3. The choice of ignoring responses y_1 and y_4 in particular was made because y_1 and y_4 have no global maxima (based on their least-squares models); their stationary points are saddle points which lie in opposite quadrants of the variable space, far from the design region.

For this case, the efficient set (based on the response models) in the factor space is a curve connecting the two individual optima. We maximize the achievement function in (24) with α_j chosen as $1/\hat{\sigma}_j$, $\beta = 0.01$, and the estimated ideal point $\hat{\phi} = [0.68\ 1.9]'$ as the reference point. While this choice of α_j is arbitrary, and while other choices are possible, it normalizes
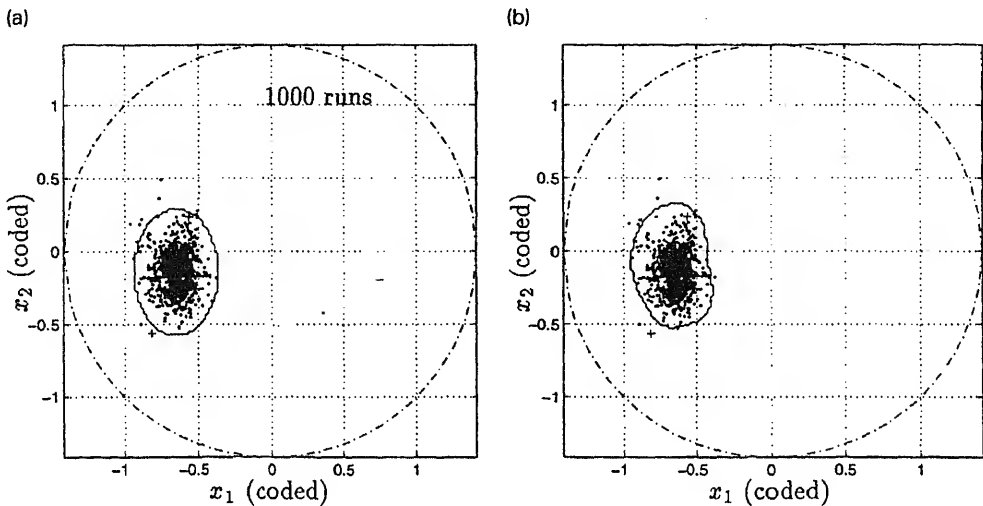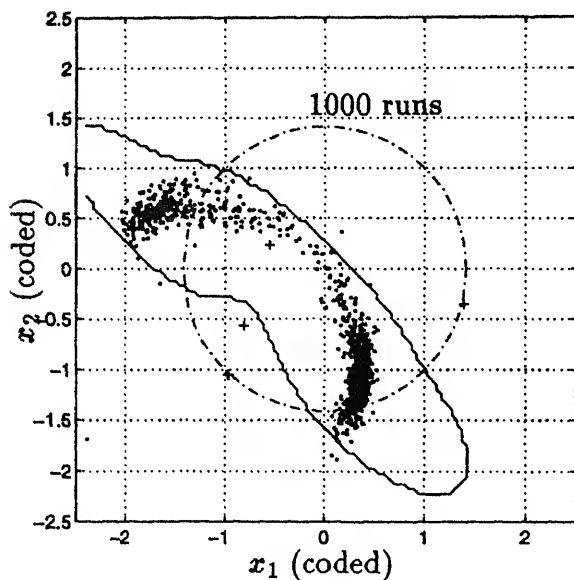
Figure 12. Distribution of the multiobjective optimum (for all four responses) for the example; the '+'s show the positions of the optima of the individual (estimated) responses. The region of significance is obtained using the statistic t_2 with up to sixth-order moments.

the achievement function against differences in the noise powers in the two responses, and also happens approximately to normalize for the different absolute scales. The estimates of σ_2^2 and σ_3^2 are 0.0005 and 0.0025 respectively. For maximization of s, a Simplex search method was used (Nelder & Mead 1965). The model optimum and its 'significance' regions obtained using the statistics of (27) and (28) (using up to fourth-order terms) are shown in figure 10. The multiobjective optimum is obtained at $(x_1^*, x_2^*) = (-0.61, -0.16)$, and the predicted response vector at this point is $(y_2, y_3) = (0.67, 1.88)$. At the individual model optima, the estimated response vectors are $(y_2, y_3) = (0.68, 1.84)$ and $(y_2, y_3) = (0.63, 1.9)$. The spread of the optimum, obtained from 1000 Monte Carlo simulations of the models, indicates the region in which the true optimum could lie assuming the quadratic model assumptions to be correct. The region enclosed by the solid line is that obtained from the Monte Carlo significance test for level 0.01; that is, the ten most extreme values were excluded from the region. For this example, the statistic in (27) seems to be adequate. The statistic t_2 is a little better in following the shape of the scatter of optima as it accounts for the sample skewness and kurtosis. In any case, these regions indicate that both responses y_2 and y_3 can be improved by lowering the variable x_1 to a (scaled) value of approximately -0.6, while not altering the variable x_2 from its current value. Since we are considering only two responses at this point, it is possible to view the attainable set Ω in the objectives space, shown in figure 11a. The scatter of the multiobjective optima is shown in figure 11b.

All four responses considered: We now apply our method to the multiobjective optimization of all four responses in this example. We again choose the reference point to be the ideal point $[2.69, 0.68, 1.9, 0.72]'$. The α_j's are again chosen to be $1/\hat{\sigma}_j$'s, and $\beta = 0.01$ ($\hat{\sigma}_1^2 = 0.0399$ and $\hat{\sigma}_4^2 = 0.0017$). The multiobjective optimum for these parameters was found at (0.37, −1.21). The scatter of the multiobjective optima obtained from the Monte
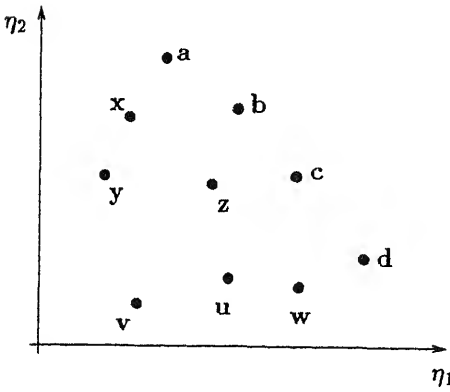
Figure 13. A two-dimensional discrete objectives set.

from the Monte Carlo significance test (for level 0.01) using the statistic t_2 with up to sixth-order terms is also shown by the solid-line curves. The scatter reveals the extent to which uncertainty in the four models can affect the position of the (multiobjective) optimum factor settings: the best factor settings could lie anywhere in a band that stretches across the experimental design region from the vicinity of $(-1, 0.5)$ to the vicinity of $(0, -1.4)$. While maximizing the achievement function in each Monte Carlo run, we have deliberately not constrained the maximum to lie inside the circle $x_1^2 + x_2^2 \leq 2$ (see the remarks following the introduction of our method). If the scatter of the global maximum of the achievement function lies mostly outside the experimental region, there is not much sense in fitting a region of significance, as a better interpretation might be to conduct more experiments in a new region outside the current design region. For fitting a region about the scatter of points in this example, the shape of the scatter demands that higher-order terms beyond the third- and fourth-order terms be included in the Edgeworth-series statistic t_2. Figure 12 shows the region obtained by including up to sixth-order terms. Since the current operating point (the design centre) also lies inside the significance region, there is not a strong case for changing the current settings for multiobjective improvement of all four responses.

5. Multiobjective optimization w.r.t. discrete factors

When quantitative factors are constrained to take only a finite number of levels, or when the factors are qualitative (or categorical), the multiobjective optimization problem reduces to a combinatorial optimization problem. Here, the assumption is that the kth factor $(k = 1, \ldots, K)$ can only take one of L_k levels: $x_k \in \{l_{k,1}, \ldots, l_{k,L_k}\}$; that is, the feasible set Ξ is finite (with cardinality $\prod_{k=1}^{K} L_k$), and so is the set Ω of feasible objectives. Using experimentation, a prediction model is obtained at all factor-level combinations. The solution of the discrete (multiobjective) optimization problem (DMOP) requires an exhaustive search in a space of cardinality $\prod_{k=1}^{K} L_k$ (which we will refer to as the size of the DMOP). However, under the frequently occurring conditions of separability of factor effects, the problem can be decomposed into problems involving much smaller search spaces.

Let the mean η_j of the jth response $(j = 1, \ldots, M)$ be related to the K factors by a relationship of the form

$$\eta_j(\mathbf{x}) = g_1^{(j)}(x_1) + g_2^{(j)}(x_2) + \cdots + g_k^{(j)}(x_K).$$

Then, it is clear that the set of x_k's, $k = 1, \ldots, K$, which optimize the respective $g_k^{(j)}$ would also optimize η_j. This approach can be extended in a straightforward way to the multiobjective optimization of all M responses, as shown by the following theorem (Song *et al* 1995).

Theorem 1. *Let the sets $B_k \subset \mathcal{R}^M$, $k = 1, \ldots, K$ and the set C, defined as*

$$C = \{\eta | \eta = g_1 + g_2 + \cdots + g_K, g_k \in B_k\},$$

have the same domination cone. Then, if $\check{g}_k$ is dominated in B_k for at least one k, $k = 1, \ldots, K$, the point $\check{\eta} = \check{g}_1 + \cdots + \check{g}_K$ is dominated in C.

Proof 1. Assume that D is the domination cone, and $\check{\eta} = \check{g}_1 + \cdots + \check{g}_K$ is a nondominated point in C. Then

$$(\check{\eta} + D) \cap C = \emptyset. \tag{29}$$

Suppose $\check{g}_k$ is dominated for some k. Then there exists a $\tilde{g}_k \in B_k$ such that $\tilde{g}_k \in \check{g}_k + D$. Since $\check{\eta} = \check{g}_1 + \cdots + \check{g}_K$, we have

$$\tilde{\eta} = \check{g}_1 + \cdots + \tilde{g}_k + \cdots + \check{g}_K \in \check{g}_l + \cdots + \check{g}_k + \cdots + \check{g}_K + D. \tag{30}$$

Thus $\tilde{\eta} \in C$ and $\tilde{\eta} \in \check{\eta} + D$, which contradicts the assumption that $\check{\eta}$ is nondominated.

The applicability of the above theorem to the discrete factors case can greatly reduce the complexity of the optimization problem, for it enables decomposition of the multiobjective optimization problem into two stages: the first stage consists of K multiobjective optimization problems of sizes $L_1, L_2, \ldots, L_K$: (multiobjective) optimize $\bar{g}_k$ with respect to x_k. The nondominated sets B_k obtained from the first stage can be used to construct the search space C for the second stage whose cardinality can be much smaller than $\prod_{k=1}^K L_k$. For example, if an experiment involves six three-level factors, the search space contains $3^6 = 729$ combinations. If the low-level (first-stage) optimization problems (each of size 3) reduce the size of the nondominated sets even by one for each factor, the size of the search space for the second stage would be reduced to $2^6 = 64$. Thus, the cost of the high-level optimization would be reduced ten-fold. If the effects of some factors do interact, the decomposition can still be done with respect to those factors that do not interact, and some reduction in complexity attained.

Search for the nondominated set: An explicit enumeration technique called the *technique of dominate approximations* (TDA) (Majchrzak 1989) can be used to obtain the nondominated set from a discrete space Ω. The technique can be illustrated using figure 13. In iteration one, the method maximizes the performance measure η_1 to obtain the point $\mathbf{d}$, and then rejects all points dominated by $\mathbf{d}$ to generate a dominated approximation $\Omega_1 = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{x}, \mathbf{y}, \mathbf{z}\}$; in the second iteration, the method maximizes component η_2 (excluding $\mathbf{d}$) to obtain $\mathbf{a}$, and then rejects all points dominated by $\mathbf{a}$ to generate another

Table 4.   Orthogonal design (L_{18}) and summarized data.

	Factors							Responses			
	1	2	3	4	5 ·	6	7	*L*	*W*	*F*	*S*
1	1	1	1	1	1	1	1	−8.6	−12.2	−2.2	1.0
2	1	1	2	2	2	2	2	−10.3	−9.2	−5.7	1.0
3	1	1	3	3	3	3	3	−15.9	−14.0	−8.5	1.33
4	1	2	1	1	2	2	3	−2.2	−7.2	−1.1	1.0
5	1	2	2	2	3	3	1	−13.0	−13.6	−7.4	1.0
6	1	2	3	3	1	1	2	−10.2	−17.5	−14.7	1.33
7	1	3	1	2	1	3	2	−13.3	−16.3	−7.7	1.5
8	1	3	2	3	2	1	3	−14.3	−16.3	−10.4	1.0
9	1	3	3	1	3	2	1	−16.9	−18.2	−14.2	2.33
10	2	1	1	3	3	2	2	−12.0	−11.7	−2.5	1.0
11	2	1	2	1	1	3	3	−15.8	−18.2	−11.1	1.0
12	2	1	3	2	2	1	1	−17.1	−8.4	−12.2	1.67
13	2	2	1	2	3	1	3	−18.1	−13.7	−9.7	3.0
14	2	2	2	3	1	2	1	−5.6	−15.1	−7.8	1.33
15	2	2	3	1	2	3	2	−16.1	−15.1	−11.8	3.0
16	2	3	1	3	2	3	1	−11.4	−16.7	−12.5	1.83
17	2	3	2	1	3	1	2	−12.1	−16.3	−11.0	1.33
18	2	3	3	2	1	2	3	−6.3	−20.9	−12.5	2.0

Factors 1: Injection pressure; 2: Injection speed; 3: Mould temperature; 4: Melt temperature; 5: Holding pressure; 6: Cool time; 7: Hold time
Responses *L*: Length SNR; *W*: Width SNR; *F*: Flatness SNR; *S*: Surface quality

dominated approximation $\Omega_2 = \{a, b, c, d, z\}$. The process continues until all dominated points are eliminated.

The above arguments are given with respect to perfectly known metric functions $\eta(x)$. When experimentation is used, prediction models for these relationships are obtained and used in place of the true relationships. To identify the significant factor-effects, an analysis of variance (ANOVA) is used. With the use of orthogonal arrays, an analysis of means may also be used for easy computation of factor effects. The problem of dealing with uncertainty in the models in the discrete-factors case would require a different formulation from that for the continuous-factors case of the previous section. In the discrete-factors case, there would obviously be no region of significance about the estimated multiobjective optimum design, but rather a set of probable designs with an associated discrete probability distribution. This problem will be addressed in future research.

5.1   *Example*

An experiment was conducted (Greenall 1989), for the optimization of a manufacturing process for injection-moulded plastic housings. The experiment involved seven factors: injection pressure, injection speed, mould temperature, melt temperature, holding pressure, cool time, and hold time; the first factor was studied at two levels, and the rest of the factors at three levels each. Four response variables were considered; overall housing length (L), overall housing width (W), flatness (F), and surface quality (S). An $L_{18}(2^1 3^7)$ orthogonal array was used. Table 4 shows the orthogonal design and the summarized data for the four responses; the first three responses have been summarized into signal-to-noise ratios,

Table 5. Results of low-level optimization.

Factors	Levels	Mean effects			
		L	*W*	*F*	*S*
Injection pressure	1	−11.63	−13.83	−7.99	1.28
	2	−12.7	−15.1	−10.1	1.80
Injection speed	1	−13.28	−12.28	−7.03	1.17
	2	−10.87	−13.70	−8.75	1.78
Mould temperature	1	−10.93	−12.97	−5.95	1.56
	3	−13.75	−15.68	−12.32	1.94
Melt temperature	1	−11.95	−14.53	−8.57	1.61
	2	−13.02	−13.68	−9.20	1.70
	3	−11.57	−15.22	−9.40	1.30
Hold pressure	1	−9.97	−16.70	−9.33	1.36
	2	−11.90	−12.15	−8.95	1.58
	3	−14.67	−14.58	−8.88	1.67
Cool time	1	−13.40	−14.07	−10.03	1.56
	2	−8.88	−13.72	−7.30	1.44
	3	−14.25	−15.65	−9.83	1.61
Hold time	1	−12.10	−14.03	−9.38	1.53
	2	−12.33	−14.35	−8.90	1.53
	3	−12.10	−15.05	−8.88	1.56

while surface quality is an average of three independent assessments on a scale of 1 to 3 (see Greenall 1989, for the complete data and analysis).

The ANOVA carried out by Greenall (1989) on the signal-to-noise ratios showed that the additivity assumptions are fairly well satisfied; so, the two-step approach presented above can be applied.

The mean responses can be computed using an analysis of means to obtain a prediction equation of the form (11). From table 4, it can be seen that a total of $2 \times 3^6 = 1458$ combinations (and, therefore, 1458 performance predictions) of different factor-levels are possible. A low-level optimization eliminates dominated points for each factor based on the mean effects (the results are shown in table 5). From table 5, it can be seen that the number of combinations from the remaining points is $2^3 \times 3^4 = 648$. The high-level optimization performed on the set of 648 reduces combinations, based on the predictions of performance measures using (11), resulted in 135 nondominated solutions. The final choice of one operating condition should be made by the design engineer. In practice, the design engineer can use the reference point approach and his/her preferences to find the most preferred solution from this nondominated set in an interactive manner. Let r be a reference point specified by the user (decision-maker). A simple strategy would be to find the closest nondominated point q, from the already determined set of nondominated solutions (denoted $P_N \subset \Omega$), to r such that $q = \arg\{\min|q - r|, q \in P_N\}$.

For illustrative purposes, each of the seven recommended designs obtained (Greenall 1989) was chosen as a reference point and the nondominated solution closest to it obtained. The results are shown in table 6. ref$_i$ ($i = 1, \ldots, 7$) in table 6 denotes the ith solution in Greenall (1989), and sol$_i$ denotes the Pareto-optimal solution found by our algorithm

Table 6. Final results of optimization.

	Predicted performance measures				
	L	W	F	S	Levels
ref_1	−8.91	−6.12	−0.99	1.03	1 1 1 2 2 2 2
sol_1	−8.91	−6.12	−0.99	1.03	1 1 1 2 2 2 2
ref_2	−11.7	−8.83	−7.36	1.42	1 1 3 2 2 2 2*
sol_2	−10.8	−7.57	−5.92	1.75	1 2 1 2 2 1 1
ref_3	−10.0	−7.40	−3.12	1.55	2 1 1 2 2 2 2
sol_3	−10.0	−7.40	−3.12	1.55	2 1 1 2 2 2 2
ref_4	−12.8	−10.1	−9.49	1.94	2 1 3 2 2 2 2*
sol_4	−11.8	−8.86	−8.06	2.27	2 2 1 2 2 1 1
ref_5	−9.83	−7.93	−3.94	0.59	1 1 2 2 2 2 2*
sol_5	−9.77	−7.08	−3.61	1.54	2 1 1 2 2 2 1
ref_6	−10.9	−9.22	−6.07	1.10	2 1 2 2 2 2 2*
sol_6	−10.8	−7.57	−5.92	1.75	1 2 1 2 2 1 1
ref_7	−10.4	−11.5	−11.2	2.55	2 2 3 2 2 2 2
sol_7	−10.4	−11.5	−11.2	2.55	2 2 3 2 2 2 2

*Dominated solutions in Greenall (1989).

using ref_i as the reference point. It can be seen that four of the original solutions, marked by *, are dominated by solutions found by our method. In other words, four of the original solutions in Greenall (1989) are not Pareto-optimal.

It is clear that an *ad hoc* approach, such as that used by Greenall (1989), can run into difficulties, if the search space of the problem is very large. The two-step approach, on the other hand, successfully finds all 135 nondominated solutions and provides a simple method for the engineer to specify preferences. The reference point itself may or may not be attainable; one can enter *any point* that reflects one's preferences, and the algorithm would always find the 'closest' nondominated point. In the case when little knowledge or preference information is available for the problem (as would happen during the initial stage of design and optimization), one can use the *ideal point* as the reference point.

6. Summary

This article has reviewed the use of experimentation to determine the best operating points for a manufacturing process, or the best design for a product's parameters, so as to optimize one or more quality criteria. After illustrating the steps in robust design used to lower variability, it specifically examined the optimization problems arising when several models for quality characteristics are estimated from experimental data. In the case of continuous factors, a new approach was discussed for dealing with the uncertainty associated with the use of response surface models for the quality metrics. This approach prevents faulty inferences from the optimization step and gives the designer or process engineer a means for determining whether to conduct further experiments or to accept the optimization results. In the case of discrete factors, an efficient search technique for the multiobjective optimal

sign was presented. The problem of developing a scheme for dealing with uncertainty
the discrete-factors case is a potential subject for future research.

ferences

derson T W 1984 *An introduction to multivariate statistical analysis* (New York: John Wiley & Sons)

rnard G A 1963 Discussion of Bartlett's paper: The spectral analysis of point processes. *J. R. Stat. Soc.* B25: 294–296

ndell A, Disney J, Pridmore W A (eds) 1989 *Taguchi methods: Applications in world industry* (Kempston, Bedford: IFS; Berlin, New York: Springer Verlag)

x G E P 1988 Signal-to-noise ratios, performance criteria, and transformations. *Technometrics* 30: 1–17

x G E P, Draper N R 1987 *Empirical model-building and response surfaces* (New York: John Wiley & Sons)

x G E P, Hunter W G, Hunter J S 1978 *Statistics for experimenters, An introduction to design, data analysis, and model building* (New York: John Wiley & Sons)

za L A, Ciaccia T J, Gatenby D A, Muise R W, Ng K K, Yanizeski G M 1994 Achieving robust design through customer satisfaction. *AT&T Tech. J.* 73: 48–58

ankong V, Haimes Y 1983 *Multiobjective decision making: Theory and methodology* (New York: North Holland)

eenall R 1989 A Taguchi optimization of the manufacturing process for an injection molded housing. In *Taguchi methods: Applications in World Industry* (eds) A Bendell, J Disney, W A Pridmore (Kempston, Bedford: IFS; Berlin, New York: Springer Verlag)

rry M J 1988 *The nature of six sigma quality* (Schaumburg, Illinois: Motorola)

uri A I, Conlon M 1981 Simultaneous optimization of multiple responses represented by polynomial regression functions. *Technometrics* 23: 363–375

lassa J E 1994 *Series approximation methods in statistics* (New York: Springer-Verlag)

ón R V, Shoemaker A C, Kacker R N 1987 Performance measures independent of adjustment. *Technometrics* 29: 253–265

wandowski A, Kreglewski T, Rogowski T, Wierzbicki A P 1989 Decision support systems of DIDAS family (Dynamic Interactive Decision Analysis & Support). In *Aspiration based decision support systems: Theory, software, and applications* (eds) A Lewandowski, A P Wierzbicki (New York: Springer Verlag)

ajchrzak J 1989 A methodological guide to the decision support system DISCRET for discrete alternatives problems. In *Aspiration based decision support systems* (eds) A Lewandowski, A P Wierzbicki (New York: Springer Verlag)

athur A, Pattipati K R 1995 Multiobjective optimization using regression models. *J. Quality Technol.* (submitted)

athur A, Pattipati K R 1996 Multiobjective optimization of process parameters using regression models. *Proceedings of CIMAT '96, International Conference on Computer Integrated Manufacturing and Automation Technology*, Grenoble, France

ontgomery D C 1991a *Introduction to statistical quality control* (New York: John Wiley & Sons)

ontgomery D C 1991b *Design and analysis of experiments* (New York: John Wiley & Sons)

ir V N 1992 Taguchi's parameter design: A panel discussion. *Technometrics* 34: 127–161

elder J A, Mead R 1965 A simplex method for function minimization. *Comput. J.* 7: 308–313

adke M D 1989 *Quality engineering using robust design* (Englewood Cliffs, NJ: Prentice Hall)

Song A, Mathur A, Pattipati K R 1995 Design of process parameters using robust design techniques and multiple criteria optimization. *IEEE Trans. Syst. Man, Cybern.* 25: 1437–1446

Steuer R E 1989 *Multiple criteria optimization: Theory, computation and applications* (Malabar, FL: Robert E Krieger)

Taguchi G 1987 *System of experimental design* (White Plains, NY: Kraus International) vols. 1 & 2

Wierzbicki A P 1980 The use of reference objectives in multiobjective optimization. In: *Multiple criteria decision making, theory and applications* (eds) Fandel, Gal (New York: Springer Verlag)

Zeleny M 1982 *Multiple criteria decision making* (New York: McGraw Hill)

Managing configurable products in the computer industry: Planning and coordination issues

RAMESH SRINIVASAN[1] and JAYASHANKAR M SWAMINATHAN[2]

[1]IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA
[2]Walter A Haas School of Business, University of California, California, Berkeley, CA 94720, USA
e-mail: sramesh@watson.ibm.com; msj@haas.berkeley.edu

Abstract. With the increase in the drive towards greater customer satisfaction, there has been a proliferation in the variety of products offered by manufacturers. Variety provides considerable choice to customers and encourages manufacturers to modularize their product lines. However, associated with a broad product line are complexities related to design, forecasting, coordination and operations. In this paper, we discuss some of these important issues.

Keywords. Product variety; feature-based product lines; operations management; forecasting; planning; computer industry.

1. Introduction

In a global economy, manufacturers are driven towards offering broader product lines as they have to cater to the needs of customers with very different requirements. This has led manufacturers to offer customizable products in different configurations utilizing common building blocks also known as feature-based product lines. Although offering more variety in the product line increases the number of market segments that are covered (Bagozzi 1986), it also increases the complexity of the manufacturing process, thereby increasing operational costs (Abeggelen & Stalk 1985). Increase in variety also increases the number of components and sub-assemblies that are utilized. For example, General Motors Corporation offered 131 different rear-axle assemblies with the intent of providing variety in pickup trucks (Fonte 1994). Such increases in variety lead to greater complexities in forecasting, parts planning, final assembly and delivery of products.

In a feature-based product line, each product is defined in terms of configurations which are built from a feature-set offered by the manufacturer. This provides customers with the ability to have their choice of features while defining the product configuration. In addition, intense competition between manufacturers has made the time to deliver, or responsiveness, an important factor in determining their success. As a result, customers have more choice in terms of product features and have greater control over lead times that they can expect

keeping costs under control they have to deal with uncertainty in customer demands and respond quickly once customer demands are known. Since the number of configurations demanded by customers is enormous, there are additional challenges to the manufacturer in demand forecasting and production planning.

In order to overcome these challenges, manufacturers are adopting several new techniques. These include incorporation of operational constraints during design of the product, delaying differentiation of products during assembly, integrating information within the organization and across the supply chain, clustering products into product families, exploiting commonality in components and manufacturing processes and using better decision-support tools for forecasting and parts planning. For example, Ford Motors has started designing cars for a global market while keeping in mind the various options that it may need to provide to different customer zones (Treece *et al* 1995). Hewlett Packard and IBM are delaying product differentiation in order to provide variety in their product line while keeping costs under control (Swaminathan & Tayur 1996). Toyota has been clustering products into families to exploit commonality in order to develop cost effective designs (Gupta & Krishnan 1995). While these techniques have shown considerable success in particular cases where they were employed, in a broader perspective it is essential to understand challenges that may arise when variety in the product line is increased.

In this paper our primary focus is on the computer industry and we highlight some of the important issues related to forecasting, parts planning, final assembly and distribution that arise as a result of increase in product variety. The rest of the paper is as follows. In § 2, we describe feature-based product structure in greater detail with an example. In § 3, we describe issues related to forecasting. In § 4, we describe some of the issues related to parts planning, final assembly and interplant coordination and in § 5 we briefly describe other issues related to managing feature-based product lines and provide our concluding remarks.

2. Feature-based product lines

In this section, we introduce the concept of a feature-based product line through an example from the personal computer industry. Traditionally, in the personal computer industry manufacturers have been offering predefined computer models to customers. Over time as a response to customer requirements the number of models grew enormously. For example, at the beginning of this decade IBM offered a few hundred models of personal computer. Such large numbers of predefined end-products resulted in complexities in forecasting and planning and inefficiencies in operations. This led computer manufacturers to experiment with new product offering strategies. One such strategy is to offer a feature-based product line. As opposed to offering numerous predefined models to customers, a feature-based product line permits customers to define their choice of configuration based on hardware and software features provided by the manufacturer. In general, a feature is a subassembly or subsystem to which a function can be ascribed from a customer perspective. Examples of features are: a graphics hardware kit, token-ring card, 1 GB direct-access storage device etc.

complete systems

peripherals, subsystems

sub-assemblies, parts
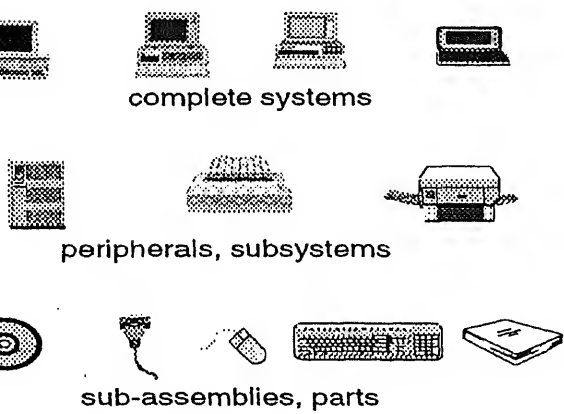
Figure 1. Customer-orderable items.

A customer's order can be made up of a collection of entities drawn from system units, subsystems, peripheral, other features and parts (see figure 1). A system unit is defined by a customer's choice of feature *categories* such as processor, hard-drive, memory, power-supply, communication and multi-media features (refer to figure 2). Certain rules are observed in the specification of a system unit. For example, a customer may not be allowed to choose a low speed processor and a very high capacity hard-drive. A feature-based approach provides adequate flexibility in defining a system unit. For example, a customer may not buy a hard-drive with a system unit, but may have it installed separately. Some features like power supply and processor are mandatory in defining a system unit while others like memory are optional. In each feature category, a customer makes a specific choice from the list of options provided. For example, there are hundreds of options available to choose to satisfy communications and multi-media capability requirements.

Each option in every feature category has a unique bill-of-materials graph as shown in figure 3. For example, part A is an assembly that is made up of 1 unit of part J, 2 units of part K and 1 unit of part L. Similar to a conventional bill-of-materials (BOM), the BOM for a feature-based product line has the characteristics of substitution among parts and effectivity windows for parts, two important factors that add complexity to planning. A substitute part is one which can be used when the primary part is unavailable. There may be more than one substitute part for a primary part and these may be primary parts in a

					system unit
6	P75	P100	P133		processor
0M	800M	1.08G	1.6G		hard-drive
M	16M	24M	40M		memory
	PS1	PS2	PS3		power-supply
	CF1	CF2	...	CF20	communications feature
	GF1	GF2	...	GF10	graphics option
	VF1	VF2	...	VF10	video option
	SF1	SF2	...	SF10	sound option

Figure 2. Flexible product structure
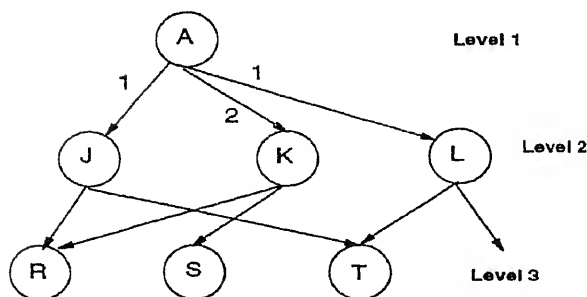
Figure 3. Bill-of-materials.

different section of the bill-of-materials structure. An effectivity window associated with a part specifies the time window during which this part can be used in manufacturing. Outside this time window, other parts specified in the bill-of-materials definition should be used. An effectivity window (defined later) for a part is usually introduced due to considerations related to engineering changes and change of supplier.

Feature-based product offering also exists in mid-range and main-frame product lines. In such product lines, a major source of demand for computer hardware and software is the customer's need to upgrade to higher capacity or the latest technology. The demand for upgrades can be greater than fifty percent and as high as ninety percent. This results in the need to offer subsystems and other modular enhancements to existing systems.

3. Implications for requirements forecasting

Plenty of research has been done on forecasting of production and inventory requirements for manufacturing industries. A good review of forecasting techniques can be found in Fildes & Beard (1992). The authors point out that serious gaps exist in the knowledge necessary to design an effective forecasting system for a production enterprise. Two key points are made in this regard. First, an effective approach to forecast demand for a product has to utilize information specific to the product. Secondly, information on which a forecast is based should include data in addition to time-series history, such as orders, marketing plans and product life-cycle. We find these conclusions to be very appropriate in the case of feature-based products.

Increased variety in product offering has important effects on product development, product upgrade possibilities and customer orders. In the following subsections, these effects and their consequences on demand forecasting are described.

3.1 *Product transition*

The product development function in a corporation has the responsibility to develop profitable products that closely match customer requirements. This is done based on market research and leads to strategies for introduction of new products and withdrawal of old products, often referred to as product transition. The function and form of new products as well as the timing of introduction are some of the decisions made and conveyed to

e forecasting and business planning units. Since the development of a product involves
ordination and completion of hundreds of activities, the actual product content realized
d the month or quarter of a year when a product becomes available for volume production
e subject to a high degree of uncertainty. Increase in product variety makes each product
velopment project very complex, thereby aggravating the difficulty in estimating the
sulting product content and time of volume production. This results in inaccuracy of
recast for existing products since time and extent of cannibalization of their demand by
w products are subject to uncertainty. In addition, sustaining existing products beyond
eviously planned months or quarters could require new parts suppliers or an increase in
ant capacity, which may be difficult to manage.

These difficulties resulting from uncertainties in product development can be handled
quantitative models. The product development project should be explicitly represented
a stochastic process made up of several parallel and sequential activities. On analyzing
ch a process, different possible completion dates for the product development activity
d their probabilities can be derived. Several demand forecasts for the new product and
isting product can then be predicated by these completion dates.

2 *Increased upgrading possibilities*

s indicated earlier, a significant portion of demand for mid-range and main-frame systems
ises from the customer's need to upgrade existing hardware capacity or upgrade to new
chnology. This results in demand for subsystems, e.g. storage systems and processors,
addition to what may be demanded with the main computer system. Upgrading also im-
ies the need for compatibility of subsystems with existing hardware. This often results
additional parts which constitute an upgrade kit. The accuracy of forecasting demand
r upgrades can benefit from approaches that consider current installed base of computers
the market place and the upgrading possibilities for the installed base of computers. In
se of high-end or main-frame systems, where the volume is low, projection of demand
r upgrades could be performed by considering each existing customer individually and
entifying upgrade possibilities depending on the customer's current product configura-
on. Since there may be several upgrade possibilities for each customer, a probabilistic
odel of upgrade possibilities is appropriate. When the customer population is large, as in
id-range computers and high-end workstations, the customer base should be segmented
cording to existing configurations before projecting demand. It is to be noted that even in
traditional product offering, customers could upgrade their hardware. However, such an
pgrade is usually from one predefined end-product to another. As a result the complexities
forecasting mentioned above are not prevalent.

3 *Customer requirements*

aditionally, a few models were offered to end-customers and forecasters could focus on
ojecting demand for each of them. Each model represented an adequately differentiated
nctionality and as a result, segmentation of customer population by functionality resulted
disjoint sets. However, in the case of feature-based product lines the offering is described

referred to as *common building blocks* (figure 2). With the usage of building-blocks in product offerings, segmentation of customer population by functionality is not precise. At the same time the number of final product configurations is too huge to develop individual forecasts with high accuracy. This has led to an alternate strategy in firms such as IBM PC Company where forecasts for individual building-blocks are generated instead of forecasts for customer demanded configurations. This process is well suited for features such as microprocessors and disk-drives which are included in most customer orders. For example, in the personal computer industry forecasters use the demand for individual commodities such as Intel's 486 processors and 4 M DRAM chips. Both end-product forecasting and building-block forecasting have their own advantages. The advantages of end-product forecasting are the following:

(1) It eases mapping to end-customer population segmentation;
(2) Characterization of demand dependencies among subsystems and sub-assemblies is easier;
(3) It eases the financial planning and monitoring at assembly plants who perform their calculations at the end-product level;
(4) It facilitates predicting customer service at the end-product level for a pre-specified safety stock allocation.

On the other hand, building-block forecasting has its own advantages which are the following:

(1) The forecasts at the building-block level are more reliable because they are aggregated across all the products;
(2) Industry level forecasts are easily available for commodities or building-blocks.

Approaches that combine the advantages of end-product forecasting and commodity forecasting do not exist and need to be developed. This is particularly essential in businesses which sell whole-systems, subsystems, sub-assemblies and parts.

4. Operational challenges

An increase in product variety and customization leads to a proliferation of components and sub-assemblies as well as increases complexity in the manufacturing process. Empirical research indicates that the degree to which the above factors influence costs, depends on the manufacturing process and techniques utilized. For example, Kekre & Srinivasan (1990) and Fisher *et al* (1993) indicate that broader product lines do not increase manufacturing cost significantly. On the other hand, MacDuffie *et al* (1996) and Banker *et al* (1990) find that product variety and complexity have a significant impact on operational costs. Kekre & Srinivasan (1990) attribute the difference to adoption of better manufacturing techniques including set-up reduction, just-in-time manufacturing, increasing commonality in the product line and providing operational flexibility.

In order to make their operations competitive, manufacturers need to pay closer attention to issues related to parts planning, final assembly and distribution. We highlight some of these issues in the following subsections.

4.1 *Parts planning*

Parts planning deals with developing detailed production and procurement plans for all the components and sub-assemblies that go into the final product. The parts-planning problem in a conventional product line, where products are pre-defined by components has been addressed by a number of researchers in academia and industry. When demands for end-products are deterministic, the problem of determining which products and how many of such products to build given constraints on component supply has been formulated as a linear program (Dietrich *et al* 1995). When demands are stochastic, the problem has been addressed by formulating it as a stochastic program and solving it by different techniques. Exploiting the commonality of components across assemblies through risk-pooling is an important feature in the stochastic problem. Industrial size problems have been addressed by the solution approach described by (Srinivasan *et al* 1992), where heuristic solutions for the stochastic problem have been proposed.

In the case of feature-based products, it will not be known *a priori* which configurations the customers will order. The number of feasible configurations could be too huge for demand to be specified for each configuration separately. This issue exists in the case of deterministic demands as well as stochastic demands. In the deterministic case the demand may be specified by a single number, *total volume*, for the entire product line. In addition, a ratio will be specified for each feature in the product line; the demand for the feature is then obtained by multiplying the total volume by the ratio. For example, one might have information that a hard disk of type A is used in 60% of orders for machine type 1111. This information in itself is not adequate because a hard disk of type A may be ordered with a higher likelihood if the product configuration has a processor of type X. Since such a specification does not explicitly show any interdependencies which may exist among features, this demand representation is incomplete. Consequently, additional assumptions are needed to determine parts requirements. Approaches to perform parts planning for feature-based product lines in deterministic and stochastic contexts are currently being pursued at IBM Research.

4.2 *Assembly planning*

Manufacturers face a tough challenge in providing a large variety to customers under conditions where demand is uncertain and a quick response is needed once it is known. Some of the leading manufacturers in the computer industry including Hewlett-Packard and IBM are trying to delay differentiation of the final product. Delayed differentiation (also called postponement) involves storing inventory in semi-finished forms and finishing the product configuration by adding components once demand is known. For example, HP decided to store inventory in the form of partially configured printers and add power supply and documentation at local distribution centres in Europe. This led to a great deal of reduction in inventory costs mainly due to risk pooling effects (Lee & Billington 1993).

customer orders for final configurations were known with certainty, additional components were added to the vanilla box and the product was shipped to the customer. An additional constraint here was that the product had to be delivered to the customer within a certain time limit from the time the order was placed. In such a situation, it was a challenging task to determine (a) how to allocate the assembly capacity between assembling vanilla boxes and final products; (b) how many types of vanilla boxes to keep (how many different partial configurations), what features to include in them and how much inventory to build each period; (c) how to allocate the existing vanilla boxes to satisfy current demand for final products. Swaminathan & Tayur (1996) provide an algorithm that can solve industry size problem within reasonable time. Their study shows that vanilla assembly process performs very well when product demands are negatively correlated and there is a high degree of commonality in the product line. Delaying differentiation at the final assembly stage helps in improving the performance of the manufacturer as compared to both assemble-to-order (all the assembly operations are done only after demand is known) and make-to-stock environments under the above conditions.

An important concern while adopting delayed differentiation is related to the correlation between demands for final products as well as the correlation between the demand for features. It seems that an ideal vanilla box is made up of features that are positively correlated in order to support a set of products whose demands are negatively correlated. Another important concern while trying to adopt such a strategy relates to capacity available at the final assembly stage. If additional capacity can be acquired at reasonable cost, then it may be optimal to operate under an assemble-to-order environment. A make-to-stock environment is preferable if the number of end-configurations is limited and demands between the products are independent or positively correlated. Swaminathan & Tayur (1996) analyse only a single level of BOM for end configurations. The more detailed problem involves consideration of features and options within each of these features. Research is in progress to address such problems.

The effectiveness of providing product variety depends to a great extent on the manufacturing process. One of the prime concerns here relates to set-up reduction. If the set-up for changing from one type to another type of the product is low then one could effectively manage production of a variety of parts or products. For example, Whitney (1993) describes how set-up reduction through better design improved the effectiveness of operations at Nippon Denso while offering more types of radiators.

4.3   *Inter-plant coordination*

Coordination between different plants of a supply chain is a challenging task. Outsourcing leads to complications in terms of coordination with suppliers as discussed in the section on parts planning. In addition, production and distribution of parts, sub-assemblies and products has many interesting aspects. Cohen & Lee (1988), Cohen & Moon (1990), Newhart *et al* (1993) have discussed supply chain coordination issues under deterministic scenarios and a global optimization criterion. Lee & Billington (1993) and Pyke & Cohen (1993, 1994) consider stochastic environments and provide approximations to optimal inventory levels, reorder intervals and service levels. Arntzen *et al* (1995) develop an

Most of the above work relates to product structures where the bill of material is fixed and known in advance. As a result, final assembly, packing peripherals and adding documentation can be coordinated in advance based on the type of product. However, in a feature-based product line, it is a complicated task to plan the merger of the subsystems comprising the order with other peripherals and documentation since the contents of the orders are not known *a priori*. In addition, it is more difficult to synchronize the arrival of the product and peripherals at the customer site since additions to the product may come from other manufacturing units. Manufacturers like IBM have primarily considered two alternatives. In the first alternative the product and peripherals are merged in transit which implies that they are synchronously received at the receiving dock. Alternatively, they are brought separately to a consolidation centre where they are merged. Consolidation centres generally have a cross-docking operation where safety stocks are not stored. In the first alternative, the challenge lies in effective transfer of information between various units in the supply chain so that the order can be completed in transit. In the second alternative, the challenge lies in effective coordination of receipt of parts at the consolidation centres so that the amount of time spent in the consolidation centre is minimized. It is to be noted that the primary difference between the two alternatives is the absence of a physical location for consolidation in the first case. For a large organization, there are also concerns regarding whether to run the consolidation centres on their own or get all the distribution services from an external vendor.

In addition to consolidating the product, there are also issues related to how products should be delivered from distribution centres to retailers. Issues related to where inventory should be stored and who should bear the financial burden for it are also very important (Anupindi & Bassok 1995). As is evident from the above subsections, operational challenges faced by a manufacturer are tougher while managing a broader product line. We have only highlighted some of the issues and have neglected those related to the shop floor such as process control, line-balancing, scheduling and batching. These issues are equally important and provide their share of difficulties while managing a broader product line.

5. Conclusions

In this paper we first introduce the notion of a feature-based product line through an example from the computer industry. Subsequently we have highlighted important challenges that manufacturers have to face in the areas of planning and operations management. In particular, we discussed issues related to forecasting, product transition, upgrading products, parts planning, final assembly and interplant coordination. We also mention some of the important problems that manufacturers face for which accurate solution methodologies are not available today.

Our focus in this paper was primarily on planning and coordination issues. However, there are other issues related to offering a broad product line some of which are mentioned below:

- Pricing decisions are extremely important for managing any product line because that in most part determines the demand generated for products. Pricing decisions are difficult

in a feature based product line because of the difficulty in segmenting the market based on functionality.

- Coordination between marketing and manufacturing in order to decide on the right set and number of features to be promoted in the product line.

- Modularity and commonality in the product line offer additional challenges while trying to integrate operations aspects in the early part of the design because one has to consider the impact of changes in design of a part on the whole product line rather than a single product. This could lead to changes in design specifications for even external suppliers.

- In high-technology industries there is an additional challenge related to choosing the right design for modules because the product life-cycles are short and a design that could be compatible with more than one generation of products may benefit from economies of scale.

- Number and location of components and final assembly plants and distribution centres need to be carefully determined because of increase in product variety and customization of different products.

Increased product variety benefits customers, while at the same time helping manufacturers to adequately differentiate their products. However, in order to successfully offer product variety a manufacturer may have to effectively overcome many of the challenges identified in this paper.

References

Abeggelen J C, Stalk G 1985 *KAISHA, The Japanese Corporation* (New York: Basic Books)

Anupindi R, Bassok Y 1995 Centralization of stocks: Retailers vs. manufacturer. Working Paper, Northwestern University

Arntzen B G, Brown G G, Harrison T P, Trafton L L 1995 Global supply chain management at Digital Equipment Corporation. *Interfaces* 25: 69–93

Bagozzi R P 1986 *Principles of marketing management* (Chicago, IL: Science Research Associates)

Cohen M A, Lee H L 1988 Strategic analysis of integrated production distribution systems. *J. Oper. Res.* 36: 216–228

Cohen M A, Moon S 1990 Impact of production scale economies, manufacturing complexity and transportation costs on supply chain facility networks. *J. Manuf. Oper. Manage.* 3: 269–292

Banker R D, Datar S M, Kekre S, Mukhopadhyay T 1990 Costs of product and process complexity. In *Measures of manufacturing excellence* (ed.) R Kaplan (Boston: Harvard Business School Press)

Dietrich B, Connors D, Ervolina T, Fasano J P, Lin G, Srinivasan R, Wittrock R, Jayaraman R

des R, Beard C 1992 Forecasting systems for production and inventory control. *Int. J. Oper. Production Manage.* 12: 4–27

her M L, Jain A, MacDuffe J P 1993 Strategies for product variety: Lessons from the auto industry. Working Paper, Wharton School, University of Pennsylvania

nte W G 1994 *A de-proliferation methodology for the automotive industry.* Master's thesis, Massachusetts Institute of Technology, Leaders of Manufacturing Program

pta S, Krishnan V 1995 Product family-based assembly sequence design to advance the responsiveness-customization frontier. Working Paper, University of Texas, Austin

kre S, Srinivasan K 1990 Broader product line: A necessity to achieve success? *Manage. Sci.* 36: 1216–1231

e H L, Billington C 1993 Material management in decentralized supply chains. *J. Oper. Res.* 41: 835–847

cDuffie J P, Sethuraman K, Fisher M L 1996 Product variety and manufacturing performance: Evidence from the international automotive assembly plant study. *Manage. Sci.* 42: 350–369

whart D D, Scott K L, Vasko F J 1993 Consolidating product sizes to minimize inventory levels for a multi-stage production and distribution system. *J. Oper. Res. Soc.* 44: 637–644

ke D F, Cohen M A 1993 Performance characteristics of stochastic integrated production-distribution systems. *Eur. J. Oper. Res.* 68: 23–48

ke D F, Cohen M A 1994 Multiproduct integrated production distribution systems. *Eur. J. Oper. Res.* 74: 18–49

nivasan R, Jayaraman R, Roundy R, Tayur S 1992 Procurement of common components in a stochastic environment. IBM Research Report, RC 18580

aminathan J M, Tayur S R 1996 Managing broader product lines through delayed differentiation using vanilla boxes. Working Paper, GSIA, Carnegie Mellon University, 1995 (revised 1996)

eece J B, Kerwin K, Dawley H 1995 FORD: Alex Trotman's daring global strategy. *Business Week* (April 3): 94–104

hitney D E 1993 Nippon Denso Co Ltd: A case study of strategic product design. Working Paper, CSDL-P 3225, Cambridge, MA-02139

Recent developments in single product, discrete-time, capacitated production-inventory systems

SRIDHAR TAYUR

239, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: stayur@grobner.gsia.cmu.edu

Abstract. We present here some recent results in single-product, capacitated production-inventory systems in discrete time. The key results are: (1) structure of optimal policy for single stage systems; (2) analysing via a shortfall process; (3) using simulation to optimize; (4) an approximation using tail probabilities. We consider periodic demand, and multiple stages – serial, distribution and assembly. Related topics of re-entrant flow shops, lead time quotation and value of information are also discussed briefly.

Keywords. Multi-echelon systems; inventory; optimal policies; simulation; approximations.

1. Introduction

The motivation for this stream of research has come from problems faced by a diverse set of companies, such as IBM, Sematech, AMD, Allegheny Ludlum, GE, Proctor and Gamble, Westinghouse, Intel, American Standard and McDonald's. Smaller local (to Pittsburgh) companies such as Sintermet, Blazer Diamond, ASKO and Northside Packing have also provided several interesting issues to pursue. At the heart of many of the problems is the interaction between demand variability and non-stationarity, available production capacity, holding costs of inventory (at different locations), lead times and desired service levels. The central goal of this research stream is to understand the interactions in simple single and multiple stage settings and to provide insights and implementable solutions for managing inventories in a cost-effective manner for complex systems. The goal of this paper is to introduce in a systematic manner some recent advances in 'Discrete-time, Capacitated Production-Inventory Systems facing Stochastic Demands' and we limit ourselves to single produc! t ! setting.

1.1 *Quantitative models for supply chain management*

A modern manufacturing network, consisting of multiple manufacturing facilities and several external vendors, can be modelled as a multi-stage, capacitated, assembly system;
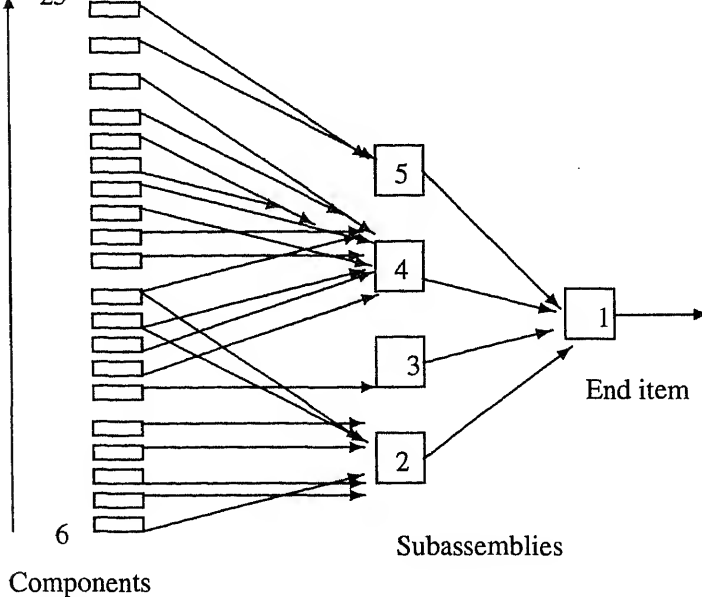
Figure 1. A typical supply chain: IBM assembly.

see figure 1 for a representation of an IBM supply chain. Until 1991, the only major result that was available in *capacitated systems* was the structure of the optimal policy for a single product, single stage system facing a stationary demand process (Federgruen & Zipkin 1986a). Even for this case ('a simple model'), no computational method was available to compute the optimal parameters for a given instance. Since then, significant progress has been made in this area. In some sense, these models form the backbone of quantitative modelling for *supply chain management*.

Among the many papers that are now available (since 1991), five papers on this topic – single-product, capacitated systems in discrete time – make the following contributions: (1) Develop a method to compute the parameters for this simple model; (2) find the optimal policy and provide a computational procedure for the case when this system faces a non-stationary (periodic) demand process (this paper generalises) (Karlin 1960; Morton 1978; Federgruen & Zipkin 1986; Zipkin 1989; Ciarallo *et al* 1994; Morton & Pentico 1995); (3) study the stability of a multi-stage capacitated system operated by a base-stock policy; (4) develop a computational method to compute good parameters for a multi-stage system operated by a base-stock policy; and (5) develop a very quick and accurate approximation method for the same problem as above. Re-entrant flow hops, multi-product systems, component commonality (and delayed differentiation) and other topics have been studied in greater detail since that time. Similarly, several papers and research! t! hemes in continuous time models are available as well.

No attempt has been made to provide a comprehensive literature review; however, most references may be obtained from the papers mentioned here. For a thorough survey of results (mainly in uncapacitated systems), see Van Houtum *et al* (1995). Other useful surveys and books include Graves *et al* (1992) and Buzzacott & Shantikumar (1993).

2 Basics

our models, we will assume *time buckets*. Each bucket can be a day or a week long r a month long) depending on the situation. These models will be called *discrete* time odels. Each unit of time will be called a *period*. This approach is appropriate at the plant d system levels. *Continuous* time models are appropriate at shop-floor and machine vels.

A *base stock* policy with order up to level z means that we produce in any period just ough to reach this target. If we cannot reach it, due to capacity limitations or lack of raw aterial, we do the best we can. An *echelon* base stock policy is exactly the same, except at all quantities (mainly inventories) considered are cumulative in order to include this age and all stages downstream (near the end product, close to the customer).

We will study these systems via a combination of analytical methods and simulation. odels are good for simple situations and to grasp concepts. To compute numbers for real orld situations, simulation is preferred. The assumptions made are more realistic and e solutions obtained are more believable. Furthermore, certain flexibility that decision akers would prefer to have is better handled by simulation. In terms of acceptance by end ers, a validated simulation has had better luck than complicated mathematical models. shortcoming of simulation as compared to mathematical models is that it takes a much nger time to find answers. What we do then in reality is use models to get rough estimates d to provide intuition to fellow team members; then choose an alternative that shows ost promise; finally, we simulate to get accurate solutions.

3 Literature survey: Papers before 1991

lark & Scarf (1960) developed a periodic review inventory control model for a serial stem without setup costs. By using a discounted cost framework, they established that order up to policy at each node is indeed optimal. Federgruen & Zipkin (1984) extended ese results further. Muckstadt *et al* (1984) conducted a computational study using the lark & Scarf (1960) model. A continuous review version of the Clark and Scarf model is udied by Debodt & Graves (1985). An in-depth analysis of an assembly structure with ly two inputs, again by using the discounted cost framework is presented by Schmidt & ahmias (1985). Rosling (1989) showed that under some initial conditions, an assembly stem can be reduced to a serial system with modified lead times so that the results of lark & Scarf (1960) may be applied to this equivalent serial system.

A model of a production and distribution network in which manufacturing is modelled / a single node is described by Cohen & Lee (1990). It also differs from much of the earlier ork in an important way in that decentralized control is assumed and the model itself a framework that combines separate models of production and distribution. A supply ain planning model that can be used to study production scale and scope economics is esented by Cohen & Moon (1990).

A model for supply chain management which assumes decentralized control at each node the manufacturing network, controlled by periodic review order up to inventory policies presented by Lee *et al* (1991). Once the service levels are set for each node, the overall lationships between cost and service can be obtained by applying this model. Although

capacity considerations are not addressed, they allow for uncertainty in the supplier lead times.

The literature on inventory control systems and production-distribution systems is extensive and hence we limit our review to the work that is closely related to the theme of this paper. Similarly, there is a vast body of literature on single location production-inventory systems (addressing many aspects of interest) that is not reviewed here.

The Clark & Scarf (1960) model and many of its extensions, including the one by Rosling (1989) analyse the model within a discounted cost framework. These results are fairly involved and further, the computational procedures are not easy to describe or program. The assembly system inventory control problem in an average cost framework is studied by Langenhoff & Zijm (1989) and Kamesam & Tayur (1993). This analysis leads to an exact decomposition of the assembly system into several single location problems. Even this decomposition is not easy to handle, but Van Houtum & Zijm (1990) describe computational approximations that lead to a simplified computational procedure.

Except for that by Federgruen & Zipkin (1986), not much work was done in capacitated systems in discrete time. In this survey paper, we will begin with their model and then describe recent developments of several extensions of this basic model. The notation we use may change between sections to remain consistent with the papers that are being summarized.

2. Single-stage, single-product models

The first progress since 1991 was the introduction of *shortfall* [1], and the connection that was made between the capacitated inventory model operated under a base stock policy and a dam model that has been studied extensively by applied probabilists.

2.1 *Basic model*

Tayur (1993) provides a method to compute the optimal policy for a basic inventory problem addressed previously by Federgruen & Zipkin (1986a). We are to determine the base stock level of a single item at a single location under periodic review, when

- the unit variable cost of purchase is c per item and there are no fixed costs;
- the holding cost (h) and stockout cost ($p > c$) are per period and per item;
- demands in successive periods are non-negative i.i.d. random variables with known distributions (labelled by d);
- there is an infinite horizon and the costs are not discounted;
- all demands that are not satisfied by stock on hand are backordered;
- there is a finite production capacity, C, in every period;
- the cost is computed on the amount of inventory or backorder at the end of each period;

---

[1] Our thanks to a referee for pointing out that this shortfall type connection was known to queueing theorists before 1991.

- we are to minimize the expected cost of holding plus penalty per period.

Recall that *inventory position* is defined as (stock on hand) + (stock on order) − (back-orders). An *order up-to* (or a base stock) policy with a critical number z is one in which the inventory position (x) is raised to z if $x < z$, and no production is done if $x \geq z$.

Unlike previous approaches, we provide a different construction of the sequence of problems that converge to the problem of interest. In particular, we do not consider finite horizon problems of the capacitated problem and then take the limit as the number of periods go to infinity. Rather, we have a sequence of uncapacitated, multi-stage, serial infinite horizon versions that converge to the desired system. We use results from uncapacitated multi-stage serial systems coupled with results in storage stochastic processes.
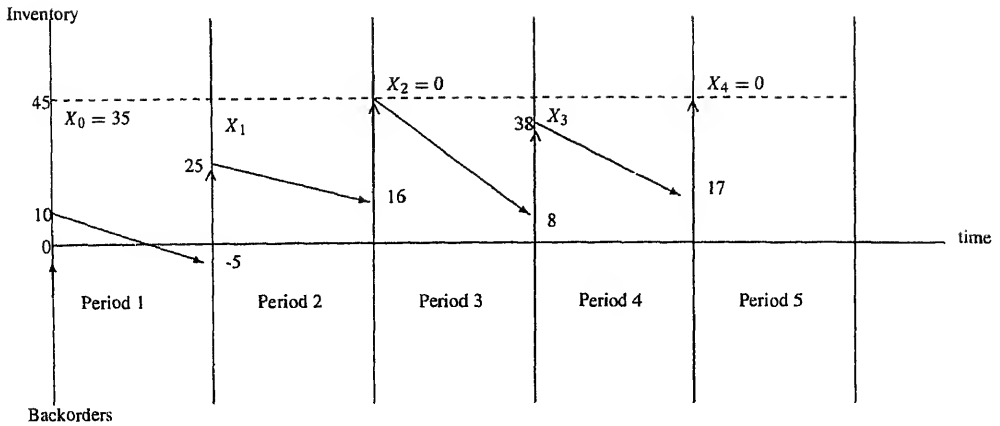
Specifically, our steps are the following. *First*, we show that the inventory model of interest is equivalent to a problem in dams. This suggests an analysis based, not on the evolution of (inventory on hand minus backorder) process but, rather, by considering a shortfall process. The shortfall is defined as the *amount on order that has not yet been produced* because of the capacity constraint. If X_n is the shortfall at the end of period n, then $X_n = \max(0, X_{n-1} + d_n - C)$, where d_n is the demand in period n.

It is important to differentiate between backorders and shortfalls: the former represents what the customer did not obtain, while the latter represents what the manufacturer could not produce because of the capacity constraint. Thus, the backorder at the end of period n equals $\max(0, X_n - z)$ where z is the order up to level. The penalty cost p is on the backorder; there is no direct penalty on shortfall. Similarly, the amount of inventory at the end of period n is $\max(0, z - X_n)$. The cost in period n, therefore, equals $p\max(0, X_n - z) + h\max(0, z - X_n)$.

Second, we show that we can replace the single-stage capacitated inventory model by constructing a specially structured uncapacitated infinite-stage inventory model: This is simply a mathematical artifact. The sequence of multi-stage problems alluded to above will converge to this infinite stage system.

2.2   Connection with a dam model

Figure 2 shows the sample path of a typical single-stage capacitated inventory system under periodic review that is operated by a base stock policy where excess demand is backlogged. The capacity (C) is 30, the order-up-to level (z) is 45, and the inventory at time 0 (I_0) is 10. Let $d_1 = 15, d_2 = 9, d_3 = 37, d_4 = 21$ be the demands in the first four periods. Figure 3 shows the sample path of a dam (see Prabhu 1965, 1980) that has an infinite height, a release capability of at most C, and an initial water level of 35. Let the rainfall in the first four periods be 15, 9, 37, and 21. The dam releases as much water as it can, and if the water level is less than C, the dam becomes empty. The equivalence of the two sample paths is obvious. Let $(a)_+$ stand for $\max(a, 0)$. If Z_n is the content of the dam in period n just after release, then it satisfies $Z_n = (Z_{n-1} + d_{n-1} - C)_+$ and if X_n is the amount on order in period n that has not yet been produced, it satisfies $X_n = (X_{n-1} + d_{n-1} - C)_+$ (a similar recursion arises in the study of a D/G/1 queue also). Note that $\{X_n, n = 1, 2, \ldots\}$ is a Markov chain. This motivates us to study the capacitated inventory system in terms of the process X_n, and provide results in terms of the steady

Inventory

$X_2 = 0$

$X_4 = 0$

45

$X_0 = 35$ X_1 38 X_3

25

16

17

10   8

0   time

-5

Period 1   Period 2   Period 3   Period 4   Period 5

Backorders

The order-up-to level is 45

The capacity is 30

Negative inventories represent backorders

X_t is the amount in period t that is on order but not yet produced

Downward sloping arrows represent demands; upward arrows denote production

Figure 2. Sample path of inventory model.

state distribution of $X = \lim_{n \to \infty} X_n$. Table 1 summarizes the equivalence between the capacitated inventory model and the dam model.

2.3 Computing the optimal base-stock level

Let $K(x)$ be the distribution of the input to the dam in any period. If $F(x)$ is the steady state distribution of the water content in the dam just after release, then the optimal value of the order up to level, z, in the capacitated inventory system satisfies

$$(F * K)(z) = p/(p + h), \tag{1}$$

where $*$ represents convolution. $F(x)$ is known for all discrete distributions of the demand, and for Erlang distributions. Intuitively, we are adding the following two independent random variables: (1) demand in a period and (2) the amount on order at the beginning of the period that has not yet been produced. The necessary and sufficient condition for the distribution $F(x)$ to exist is that the expected demand (input to the dam) in a period be less than the capacity, C. Intuitively, the result states that it is the sum of two independent random variables (demand in a period and the amount on order not yet produced at the beginning of that period) that adds up to z. Penalty p is incurred if this sum crosses (at the end of the period) z and is proportional to the excess, and a holding cost (h) is imposed if the sum is less than z and is proportional to the amount on hand at the end of the period.

Example 1. If the demand is exponentially distributed with mean rate λ ($K(u) = 1 - e^{-\lambda u}$, $u \geq 0$; $K(u) = 0$, otherwise) and the capacity is C, then the steady state distribution of the water content in the dam just after release is given by

WATER CONTENT
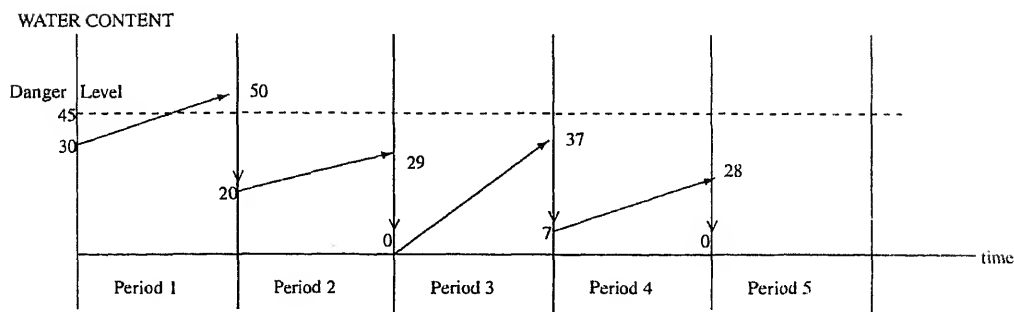
Upward arrows represent rainfall

Downward arrows represent release

One cannot release more than the water level

The danger level is simply shown to connect with the inventory model

Figure 3. Sample path of dam model.

where μ is the largest positive root of the equation

$$\mu = \lambda - \lambda e^{-(\mu+C)}.$$

Thus, the optimal order upto policy has a critical number z^*, and is obtained by solving $(F * K)(z^*) = p/(p+h)$.

Independently, the similarilty of the basic inventory model operated via a base stock policy to a D/G/1 queue is recognized by Van Houtum & Zijm (1994).

2.4 *Optimal policy for an extended model*

In the previous subsection, we computed the optimal base stock level. We now want to show that among all policies, a base stock policy is optimal. Federgruen & Zipkin (1986a) show this for the stationary case above. Kapuscinski & Tayur (1996a) provide a proof for an extension that allows for periodic demand as well as capacities by considering the following variant of the basic single-stage, single-item, discrete-time production–inventory model. The demands (stochastic) follow a periodic pattern with a period K. As before, there is a maximum production capacity (C) in any period; demands not satisfied in a period are backlogged to the next; there is holding cost (h) per unit of inventory per period and a cost of penalty (p) per unit of backlog per period. We want to find policies that minimize the

Table 1. Comparison of dam and inventory models.

Dam model	Inventory model
Maximum release	Capacity
Water content	Amount not yet produced (shortfall)
Empty dam	Order-up-to level achieved
Rainfall	Demand
Danger level crossed	Backorders

finite-horizon costs, the infinite-horizon discounted cost and the infinite-horizon average cost (respectively) of operating this stage.

For all the three cases – finite-horizon cost, discounted infinite-horizon cost and infinite-horizon average cost – we show that an order up-to (or base-stock, critical-number) policy is optimal. This extends the results of Karlin (1960b) and Zipkin (1989) for uncapacitated, non-stationary models and Federgruen & Zipkin (1986) for capacitated, stationary models. Our proof for the finite-horizon case follows standard steps. The proof of optimality for the infinite-horizon discounted case is simpler than that provided by Federgruen & Zipkin (1986) because we are able to use more recent results from Bertsekas (1987). To provide the optimality proof in the average cost case, we use the framework of Federgruen *et al* (1983), but our approach is different from that used by Federgruen & Zipkin (1986a) for the stationary case. All proofs can be found in Kapuscinski & Tayur (1996a).

2.5   *Sequence of results leading to optimal policy structure*

The following is the sequence of events at the beginning of a period: (1) some inventory or backlog exists; (2) a decision to increase the inventory is taken (limited by the production capacity); and (3) demand arrives. Holding or penalty costs are charged on the inventory after demand arrives. The notation is mostly standard. We have suppressed the time subscript in x, y and d below and these are assumed to be reals unless mentioned explicitly as integers. We will write them when necessary. We define

- x: inventory at the beginning of a period;
- y: inventory after ordering, but before demand arrives.

We assume that $E d_i < \infty$ for all period types $i = 1, \ldots, K$.

2.5a   *Finite horizon*:   As is standard practice, we begin with the finite-horizon case, which is not only the simplest situation to consider but also sets the framework for the infinite horizon cases.

Let $0 < \beta \le 1$ be the discount factor. Define (recursively) $v_n(x) =$ minimum total expected discounted cost with a time horizon of n periods. Note (in this subsection only) that *we start in period n and count downward towards period 1*, the end of the horizon. The demand in period n is one of the K period types ($i = 1, \ldots K$). We may assume that period 1 has type 1 demand and period $n = K + 1$ has type 1 demand again and so on. Thus:

$$v_0(\cdot) = 0,$$
$$v_n(x) = \inf_{(x,y)\in A} \{c(y - x) + L_n(y) + \beta E_{d_n} v_{n-1}(y - d_n)\},$$
$$\text{for } x \in R \text{ and } n \ge 1,$$

where

$$A = \{(x, y) \in R^2 | x \le y \le x + C\} = \bigcup_{x \in R} Y(x)$$

is the feasible set. For $x \in R$,

$$Y(x) = [x, x + C],$$

and for $y \in R$,

$$L_n(y) = E_{d_n}(h(y - d_n)^+ + p(y - d_n)^-).$$

We can express v_n using additional functions J_n and I_n as follows.

$$v_n(x) = -cx + I_n(x), \quad x \in R, \quad n \geq 1,$$
$$I_n(x) = \min\{J_n(y) : y \in Y(x)\}, \quad x \in R,$$
$$J_n(y) = cy + L_n(y) + \beta E v_{n-1}(y - d_n), \quad y \in R,$$

with $v_0(\cdot) = J_0(\cdot) = I_0(\cdot) = 0$.

Our first lemma generalizes theorem 2 of Federgruen & Zipkin (1986b) to the cyclic case. The proof is a direct adoption of theirs.

Lemma 1. *The set of feasible pairs* (x, y), A *(as defined above), is convex. For all* $n \in N$:

(a) *The expected sum of holding and penalty costs,* $L_n(y)$, *is convex;*

(b) J_n, v_n, *and* I_n *are convex;*

(c) $v_n \geq 0$; *and*

(d) *For* $n \geq 1$: $v_n(x_n) \to \infty$ *when* $|x_n| \to \infty$, *and if* $p > c$ *then* $J_n(y_n) \to \infty$ *when* $|y_n| \to \infty$.

Theorem 1. *Let* y_n^* *be the smallest value minimizing* J_n. *The optimal policy in period* n *is order up-to* y_n^*. *As* $\lim_{|y_n| \to \infty} J_n(y_n) = \infty$, *we have* $y_n^* < \infty$.

As a first property, we have the following.

Property 1. For any $x \in R$, $v_{nK+i}(x)$ is increasing in n.

Proof. We show by induction that $v_{m+K}(x) \geq v_m(x)$ for all x and for all m. First, $v_K(x) \geq 0 = v_0(x)$. Let $v_{K+m}(x) \geq v_m(x)$ for a certain m and all x. Then,

$$
\begin{aligned}
v_{K+m+1}(x) &= \min_{y \in Y(x)} \{c(y - x) + L_{K+m+1}(y) + \beta E v_{K+m}(y - d_{K+m+1})\} \\
&\geq \min_{y \in Y(x)} \{c(y - x) + L_{m+1}(y) + \beta E v_m(y - d_{m+1})\} \\
&= v_{m+1}(x). \quad \square
\end{aligned}
$$

Note that the convexity of functions J_n, I_n, v_n implies continuity of these functions. One-sided derivatives exist at all points. Also, two-sided derivatives exist with the exception of a denumerable set of points. Points where two-sided derivatives do not exist are generated by mass points of demands. Although derivatives do not have to be continuous, they are monotonic and bounded on any compact set. Therefore, in this paper we will define them as right-hand-side limits. We will use a prime (') to denote these derivatives.

We define the *myopic* solution to period i, $\bar{y}_i$, as the one that satisfies $c + L_i'(\bar{y}_i) = 0$. Property 2 provides a simple lower bound as in the uncapacitated case; see Zipkin (1989).

Property 2. Assume that period 1 has the minimum myopic solution; $\bar{y}_1 = \min\{\bar{y}_i :$ $i = 1, \ldots, K\}$ where $\bar{y}_i$ is the myopic solution for period i. Then: $y_n^* \geq \bar{y}_1, \forall n$.

Strictly speaking, our next result is not a finite-horizon result. However, this appears to be the most appropriate point to state it. Part (c) of the technical lemma (lemma 2) is used for the next result. We will use this property in proving property 6 in § 2.5a below and when analysing the infinite horizon average cost case in § 2.3.

Property 3. For a given $0 < \beta \leq 1$, the sequence y_n^* is bounded. That is, $\limsup\{y_n^* : n \in N\} < \infty$.

Property 4. For any finite-horizon (n periods) problem consider a policy that produces up-to z_i in period $i = 1, \ldots, n$. For all i, the cost of such a policy is convex in z_i.

2.5b *Infinite horizon: Discounted model* :   We now move to the infinite-horizon discounted case. It was natural in the finite-horizon case to label periods as time to the end of the horizon. In the infinite-horizon case, we typically start the process at some point and continue indefinitely. To make the notation more intuitive, starting from this subsection, *we will number periods in increasing order*: following period n, we have period $n + 1$.

Federgruen & Zipkin (1986b) showed the next result for the stationary, capacitated case. We provide a simpler proof as we are able to use results from Bertsekas (1987).

Theorem 2. *Let $0 < \beta < 1$ (discounted case). The optimal policy for the infinite-horizon is cyclic up-to level policy.*

Since $E(d_i) < \infty$ for $i = 1, \ldots, K$, we have the following.

Property 5. Let $w_n(x, i) := v_{nK+i}(x)$ for $i = 1, \ldots, K$, $n \in N_0 := N \cup \{0\}$, and $x \in R$. The limit, $\lim_{n \to \infty} w_n(x, i) = w(x, i)$ (where values of w are in $R \cup \{\infty\}$), exists and $w < \infty$.

$w < \infty$ does not imply that stationary up-to levels are finite. However, it is possible to prove the following (since $E(d_l) < \infty$ for $l = 1, \ldots, K$).

Property 6. The up-to levels are finite (i.e. $z_l < \infty$).

2.5c *Average cost criterion* :   This case is the most difficult one to analyse. We need some technical results before the optimality of base-stock policies can be proved.

Property 7. (Convexity) Consider the infinite-horizon case (with cyclic demands) and the class of up-to level policies, where levels z_i for $= 1, \ldots, K$ are period-type specific. For both the discounted cost and average cost criteria, the average cost is finite and convex in each of z_i's.

Let $z_{\max}$ and $z_{\min}$ be the maximum and minimum respectively among the levels for a

starting points x_{i_0} and x_{i_0}' respectively in period t_0, there exists (with probability 1) a period n such that starting from this period the two processes coincide.

Property 9. (Shortfall stability) Consider an up-to policy with a vector $z_1, \ldots, z_K$. If $\sum_{i=1}^{K} E d_i < KC$ then $E S_i < \infty$ for all i's, where $S_i = z_i - y_i$ is the shortfall in period type i.

Average cost criterion is easy to analyse when either the number of states the process can take is finite or the one-period cost function is bounded (see Bertsekas 1987). Conditions when an optimal policy exists for semi-Markovian process with average cost objective function with denumerable state space and unbounded one-period cost function are derived by Federgruen *et al* (1983). These conditions were used by Federgruen & Zipkin (1986a) to derive optimality of up-to policy for a capacitated stationary model. We extend it to a cyclical model as follows.

We first show that any policy can be dominated by a policy that requires reducing (any) backlog and not exceeding some stationary level A^*. Then we show that among such policies, the up-to policy is optimal. The main structure of our proof is based on results of Federgruen *et al* (1983), but the proof that the required conditions are satisfied is shown by a different method as compared to that by Federgruen & Zipkin (1986a).

Fact 1. If an optimal policy for the problem exists (including possibility of randomized policies), then it has the following form:

(a) for $x \leq -C$, $y = x + C$.

(b) there exists $A^* < \infty$ such that for $y \geq A^*$ for all period types it is better to produce nothing rather than take any other action.

Theorem 3. *Consider a capacitated system with cyclic discrete demands and linear ordering, penalty, and holding costs. For the average cost criterion, the cyclic up-to policy is optimal.*

Lemma 2. Consider the policy $\delta[0]$ (produce up-to 0). Let $\sum_{i=1}^{K} E d_i < KC$ and $E(d_i)^{2k+2} < \infty$ for a certain $k \geq 1$ and for all $i = 1, \ldots, K$. Consider a point process defined by points i, for which $y_i = 0$ (no backlog of previous demands). Let N be a random variable equal to time between two consecutive points of this point process. Then for any starting period-type:

(a) $E(N^k) < \infty$

(b) $E(|y_i|^k) < \infty$ and $E(|x_i|^k) < \infty$

(c) If $0 < \beta < 1$ then $E(\sum_{n=0}^{\infty} |y_n|^k \beta^n)$, $E(\sum_{n=0}^{\infty} |x_n|^k \beta^n) < \infty$ for any x_0.

(d) For $0 < \beta \leq 1$, there exists $\dot{A} \in R$, such that $J_n(0) \leq A_n$.

2.6   *Basic properties*

We show several properties of the optimal policy including the following: (1) capacity smooths the base-stock levels in a manner that is different from that due to holding costs; (2) the limit of finite horizon order up-to levels are bounded; (3) the optimal levels are higher than the minimum of the K myopic levels; (4) in an infinite-horizon average cost case, the optimal levels are lower than the maximum of the K stationary optimal levels and higher than the minimum of the K stationary optimal levels; (5) if demands are stochastically larger or the capacity is lower, the base-stock levels are higher; and (6) for $K = 2$, as the penalty cost is increased, the difference between the maximum and minimum levels is bounded by C under fairly general assumptions on demand distributions.

2.7   *Computational technique*

Exact computation of optimal levels by analytical formulas appears difficult. We provide a simulation based method using infinitesimal perturbation analysis (IPA) to find these levels. The basic idea is simple: instead of using the derivative of the expected cost in a gradient search method, we use the expected value of the sample path derivative (obtained via simulation). To validate this approach and prove the optimality result for average cost case, we derive several technical properties of base-stock policies – convexity, regeneration, coupling and stability. See Glasserman (1991) for an excellent reference and a later subsection for details of a multi-stage system analysed in this manner. A numerical study indicates that our IPA method is robust and finds solutions within a few minutes on a workstation. The steps are similar to those described later in § 3 for a serial system, and so we do not detail them here. See Kapuscinski & Tayur (1996a).

2.8   *Insights into some complex issues*

We also numerically study several issues that provide insight into the behaviour of optimal solutions. Examples of issues studied include the following: (1) what is the increase in cost if all periods are forced to have the same base-stock level? (2) what are the benefits of changing capacities in each period based on the demand type? (3) how many periods are affected by smoothing of the levels? (4) how are the above results affected by high penalty cost or high utilizations? (5) what is the relationship between service level and costs? Several of the qualitative and technical properties in our capacity setting differ from the uncapacitated non-stationary model.

   See Kapuscinski & Tayur (1996a) for extensive computational testing. The basic insights are as follows.

(1) *Basic insights.* Several basic properties of the system confirm what we expect. All other things being equal:

   - with an increased mean demand, the up-to levels increase,
   - with decreased capacity, the up-to levels increase, and
   - with increased variance of demand, the up-to levels increase.

(2) *Effect of capacity and range of demands on 'smoothing'.* By smoothing, we mean that values z_i are affected by demands and the capacity of period $j \neq i$. The difference between z_{max} and z_{min} is lower as compared to the difference between the maximum and the minimum of K independent stationary capacitated models; so some higher levels are brought down and some lower levels are lifted up. We have two types of smoothing working in opposite directions: (a) in anticipation of high demand, some levels are lifted up, and (b) in anticipation of low demand, other levels are decreased. The second type was described by Karlin (1960b) and Zipkin (1989) in the uncapacitated setting. The first one is induced by finite capacity, and the second by the holding cost.

At times there are changes in the ordering of the levels as capacity decreases. For example, the period with the minimum level may be different depending on the capacity. Furthermore, although the differences between levels do not disappear as capacity is reduced, the levels increase and the ratio between the maximum and the minimum of the levels gets closer to 1.

(3) *Constant base-stock versus optimal (period wise) base-stock.* We compare the cost of the optimal *constant* base-stock policy to the optimal cost. (Note that the cost is a convex function of this constant level and the simulation based method remains valid.) Not surprisingly, the optimal constant z lies between the minimum and the maximum optimal base stock levels When capacity is tight, the cost ratio (say R) becomes close to 1. We note that in some cases the difference $z_{max} - z_{min}$ first increases and then decreases as capacity decreases, while in other cases it only increases with capacity. Even in cases where $z_{max} - z_{min}$ of optimal policy increases as capacity decreases, R monotonically decreases to 1. One explanation is the fact that the cost function is relatively flat around the optimum (see item 8 below), and more so as the capacity becomes tighter.

The effect of variance on R is as follows. R is close to 1 in most cases. A case when the cost ratio is large has two small but different variances (example: $N(70, 1)$, $N(70, 10^2)$) in a $K = 2$ situation). With increase in variance the levels increase and the ratio of costs goes to 1. We can also show that even with significantly different means, the ratio of costs goes to 1 when at least one of the variances is high.

(4) *Changing capacity with periods according to period type.* Rather than remain constant, capacity can vary according to period type. (The cost is still a convex function of $(z_1, \ldots, z_K)$ and the optimal policy is still order up-to. The simulation-based method remains valid and recursions can be easily adapted.) Several capacity allocations for four situations each with $K = 2$ were tested: (i) both period types have low variance; (ii) period with lower mean has a high variance; (iii) period with higher mean has a high variance; (iv) both periods have high variance. We are able to test if two natural alternatives to constant capacity (with the same total capacity KC in a cycle) – (a) proportional to mean demand and (b) proportional to (mean + constant ∗ standard deviation) of demand – do well.

either. Typically the spread proportional to mean demands did well only in small-variance cases, but in these cases nearly all spreads did well. In case of high variance, a proportional (based on mean and variance) spread did much better than one based on mean alone, but surprisingly it was possible to find cases when it was optimal to allocate more to a period type with a demand having both a smaller mean and a smaller variance. We think that in these cases there is typically sufficient capacity in low demand period to raise the inventory to optimal level in higher demand period, but the effect of a surge in demand in high demand period can be repaired faster when a little bit more capacity is assigned to low demand period. This is seen very clearly in the following example: let $K = 4$ with demands wi! th! means and variances equal to $(20, 1)$, $(20, 1)$, $(90, 50)$, $(20, 1)$ respectively. We find that the optimal capacity allocation is $(22, 22, 93, 143)$. So the use of the term 'recovery capacity' seems appropriate. In all cases, allocating equal capacity to periods did very well.

(5) *Service level.* In all the experiments – constant capacity, equal base-stock levels, period-wise capacity allocation, different K's, different demand distributions, different holding and penalty costs – the type-1 service, defined as P (inventory after demand > 0), at optimality, was $p/p + h$. See Tayur (1993) for a proof of this connection between service level and costs in the single-stage capacitated stationary case, and Glasserman & Tayur (1995) for the multistage stationary capacitated case.

(6) *Increasing penalty cost.* With increased penalty the up-to levels increased. Obviously larger differences in means and variances caused relatively larger differences in up-to levels, but this effect was small over a range of penalty costs. For most part, the levels rise in parallel. Furthermore, in all our experiments, the ordering of the levels did not change as a function of p (unlike item 2 above). We also noticed that the difference between the maximum and the minimum levels at optimality was bounded by C for $K = 2$ as we increased p. Property 12 below provides an explanation.

(7) *Mean utilization vs. variance.* As penalty cost is increased, the optimal up-to levels of low variance system with high utilization increase faster than in a system with high variance and lower utilization. This indicates that the rate of increase of base-stock levels depends on both the variance and excess capacity.

(8) *Cost sensitivity.* Cost is not very sensitive to the up-to levels around the optimum. This was noticed in the stationary multi-stage system studied by Glasserman (1991) also.

3. Single-product, serial system

It has been shown that for a serial capacitated system, the optimal policy (under the cost criteria discussed above) is not base-stock in general; see Speck & Van der Waal (1991).

3.1 *A simulation-based optimization procedure: Single-stage simulation*

Let s be any base stock level; the optimal value for s is what we eventually need. The notation used is consistent with the papers referenced on this topic where detailed proofs are available.
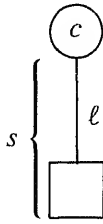
Figure 4. Illustration of a single stage.

We need notation.

$$I_n = \text{inventory} - \text{backlog in period } n;$$
$$R_n = \text{production in period } n;$$
$$\ell = \text{leadtime from production to inventory};$$
$$T_n = R_{n-1} + \cdots + R_{n-\ell}$$
$$= \text{pipeline inventory};$$
$$D_n = \text{demand in period } n;$$
$$s = \text{base-stock level};$$
$$c = \text{production capacity}.$$

We always assume that $c > 0$ and $D_n \geq 0$ for all n. Under a (modified) base-stock policy, the production level in each period is set to try to restore the inventory position $I_n + T_n - D_n$ to s. If production were uncapacitated, this would be achieved by setting $R_n = D_n$. Since, however, R_n cannot exceed c, it may take multiple periods of production to offset demand in a single period.

The system evolves as follows:

$$I_{n+1} = I_n - D_n + R_{n-\ell}, \tag{2}$$
$$R_n = c \wedge [s + D_n - (I_n + T_n)]^+, \tag{3}$$
$$T_{n+1} = T_n + R_n - R_{n-\ell}. \tag{4}$$

x^+ denotes $\max(0, x)$, x^- denotes $\max(0, -x)$ and $a \wedge b$ denotes $\min(a, b)$.

What about costs? In period n the cost will be $C_n = hI_n^+ + pI^-$. The average cost over the long term will be $(\sum_{n=1}^N C_n)/N$ for a large N. N is the number of periods simulated.

3.2 *A simulation-based optimization procedure: Single-stage gradient*

We can certainly find the gradient of expected cost by simulating one more time and changing (say, increasing) only the order up to level by 1: the difference in cost between the two simulations is the gradient. This would be too much work especially if we had several stages whose order up to levels we want to optimize. Can we get the gradient by not doing any more simulations? Yes; and here goes the basic trick.

(1) If we had inventory in period n, having started at a higher base stock level means an *increase* in holding by h.

(2) If we had backlog in period n, we will *decrease* penalty by p.

Thus, the changes are either h or $-p$. We will just average these changes. The point is we are claiming that average of changes is the change in the average. In this situation, and in several others, this claim is good.

How do we make the computer do this? In principle, we are just differentiating the recursions. The easiest way is to write the following two lines in the code to be executed in every period:

$$\text{dCost} = \text{dCost} + h, \quad \text{if } I_n > 0,$$
$$\text{dCost} = \text{dCost} - p, \quad \text{if } I_n < 0.$$

Then dividing dCost by N provides the derivative. Note that at the beginning of the simulation dCost is set to zero.

3.3   *A simulation-based optimization procedure: Serial system*

3.3a   *Operation*:   We now link multiple stages in series. There are m stages: stage 1 supplies external demands and stage i supplies components for stage $i - 1, i = 2, \ldots, m$. Stage m draws raw materials from an unlimited supply. To specify an echelon-inventory base-stock policy for the system, we let

$$s^i = \text{echelon base-stock level for stage } i.$$

Naturally, we require $s^1 \leq s^2 \leq \cdots \leq s^m$. Let the variables R_n^i, T_n^i, ℓ^i, and c^i have the same meaning as before, applied to stage i. For $i = 2, \ldots, m$, let I_n^i be the installation inventory at stage i, and let $I_n^1 = I_n$, with I_n as in § 2.1. In period n, stage i sets production to try to restore the echelon inventory position,

$$\sum_{j=1}^{i}(I_n^j + T_n^j) - D_n,$$

to its base-stock level s^i.

Two features distinguish the multi-echelon system from a single stage: production at stage $i, i < m$, is constrained by available component inventory I_n^{i+1}, as well as by the capacity limit c^i; and for $i > 1$ the amount removed in period n from the store at stage i is the downstream production level R_n^{i-1}, rather than the external demand. Thus, for stage $i = 2, \ldots, m - 1$ we have

$$I_{n+1}^i = I_n^i - R_n^{i-1} + R_{n-\ell^i}^i, \tag{5}$$

$$R_n^i = c^i \wedge \left[s^i + D_n - \sum_{j=1}^{i}(I_n^j + T_n^j)\right]^+ \wedge I_n^{i+1}, \tag{6}$$

$$T_{n+1}^i = T_n^i + R_n^i - R_{n-\ell^i}^i. \tag{7}$$

At stage m, raw materials are unlimited so the last term in (5) is absent. To subsume these special cases in (5–7), we take $R_n^0 \equiv D_n$ and $I_n^{m+1} = \infty$ for all n. To complete our specification of the model, we need initial conditions; for simplicity, we take $I_1^1 = s^1$,

$= s^i - s^{i-1}$, $i = 2, \ldots, m$, and all other variables zero. In other words, the system [st]arts with full inventory. For details see Glasserman & Tayur (1995).

Similar to the single stage case, the derivatives with respect to the base stocks can be [co]mputed.

4 *Validation of technique*

[Va]lidation of finite horizon derivatives – inventory and costs – is quite straightforward. [W]e show that (right-side) derivatives exist with probability one at a given value of s, then [ap]peal to Lipschitz continuity and finish by applying the dominated convergence theorem. [Se]e Glasserman & Tayur (1995) for details.

5 *Stability and recurrence*

[Fo]r IPA to work in the infinite horizon, several conditions have to be satisfied by the [un]derlying stochastic process. These are derived by Glasserman & Tayur (1994). When [ca]pacity limits are introduced, an ineffective policy may lead to increasingly large or-[de]r backlogs: the *stability* of the system becomes an issue. In this paper, we examine [th]e stability of a multi-echelon system in which each node has limited production ca-[pa]city and operates under a *base-stock* policy. We show that if the mean demand per [pe]riod is smaller than the capacity at every node, then inventories and backlogs are [sta]ble, having a unique stationary distribution to which they converge from all initial [sta]tes. Under i.i.d. demands we show that the system is a Harris ergodic Markov chain [an]d is thus wide-sense regenerative. Under slightly stronger conditions, inventories re-[tur]n to their target levels infinitely often, with probability one. We discuss cost implica-[tio]ns of these results, and give extensions to systems with random leadtimes and periodic [de]mands.

6 *Extensions*

[Le]t us discuss two obvious extensions.

[6]a *Assembly system*: In an assembly system, each node i requires components from [a s]et of predecessor nodes. These are assembled into stage-i finished goods. By changing [un]its, if necessary, we may assume that components from predecessor stages are assembled [in] equal quantities.

To keep the notation simple, we consider a representative example, rather than the [ge]neral case. Figure 5 depicts a three-node system in which node 1 assembles components [su]pplied by nodes 2 and 3. Node 1 feeds external demands; the other nodes draw raw [m]aterials from infinite sources. The evolution of inventory at node 1 is characterized by

$$
\begin{aligned}
I^1_{n+1} &= I^1_n - D_n + R^1_{n-\ell^1}, \\
R^1_n &= c^1 \wedge [s^1 + D_n - (I^1_n + T^1_n)]^+ \wedge I^2_n \wedge I^3_n, \\
T^1_{n+1} &= T^1_n + R^1_n - R^1_{n-\ell^1}.
\end{aligned}
\tag{8}
$$

[Th]e assembly feature is reflected in the dependence of R^1_n on I^2_n and I^3_n in (8). Nodes 2 [an]d 3 are characterized by the basic recursions (5–6), with obvious modifications to the

indexing. The only notable difference is that now I_n^2 and I_n^3 are decreased by the same production level R_n^1 each period.

3.6b   *Distribution system*

Another variant of the serial system allows intermediate stages to supply multiple lower-echelon stages, typically in a tree topology. Our results extend without difficulties to such models. For ease of exposition we describe a less general setting – a serial system in which each stage faces external demands for components in addition to internal demands from the downstream stage. A manufacturer of electronic equipment, for example, may face demands for integrated circuits, and for circuits assembled into circuit packs, along with demands for finished goods.

To characterize the operation of such a system, we need to specify how each stage allocates inventory to internal and external demand. Rather than restrict ourselves to any one policy, we describe a class of policies consistent with our results. Let I_{0n}^i be the stage-i inventory reserved for external demands at stage i in period n, and let $I_{1n}^i = I_n^i - I_{0n}^i$ be the inventory available for downstream production. Suppose stage i has base-stock levels s_0^i and s_1^i to supply external and internal demands, and let $s^i = s_0^i + s_1^i$. Denote by D_n^i the external demand at stage i in period n.

The operation of stage 1 is unchanged. At stage i, production is now set according to

$$R_n^i = c^i \wedge \left[s^i + \sum_{j=1}^{i} (D_n^j - I_n^j - T_n^j) \right]^+ \wedge I_{1n}^{i+1}.$$
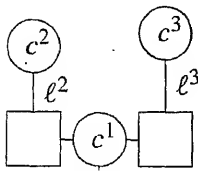
Also,

$$I_{n+1}^i = I_n^i - R_n^{i-1} - [D_n^i \wedge I_{0n}^i]^+ + R_{n-\ell^i}^i.$$

Pipeline inventory follows (5), just as before. It only remains to specify how I_{0n}^i and I_{1n}^i are determined.

A broad class of policies sets

$$I_{1n}^i = g^i \left(I_n^i, s_1^i + D_n^1 - I_{1n}^i - \left(\sum_{j=1}^{i-1} I_n^j + T_n^j \right), s_0^i + D_n^i - I_{0n}^i \right)$$

and $I_{0n}^i = I_n^i - I_{1n}^i$, for some function g^i; in other words, the inventory reserved for internal production is a function of the available inventory and of the shortfalls in meeting internal and external demands. The choice of g^i determines the particular policy.

We briefly describe two types of policies. The first type gives strict priority to internal or external demands, allocating as much inventory as needed to the high-priority demand and using any excess for the other. The second type attempts to balance shortfalls, allocating inventory to minimize the resulting difference between inventories and base-stock levels. This type of policy implements relative priorities through the choice of s_0^i and s_1^i. The assumption that the system can, in fact, be balanced in each period is essential to many analytical approaches; our setting, of course, does not require this. For both types of policies just described, the particular functions g^i are tedious to write out but they are sufficiently smooth to allow derivative calculations.

3.7 *An approximation using tail probabilities*

The IPA method could be computationally prohibitive for large problems, or for problems where the starting solution for the simulation is far way from its eventual optimal. So a quick (and accurate!) approximation is presented by Glasserman & Tayur (1996): Recall that our objective is to find base-stock levels that approximately minimize holding and backorder costs. The key step in our procedure approximates the distribution of echelon inventory by a sum of exponentials; the parameters of the exponentials are chosen to match asymptotically exact expressions. The computational requirements of the method are minimal. In a test bed of 72 problems, each with five production stages, the average relative error for our approximate optimization procedure is 1.9%.

Much of the technical development of this approximation is based on the work by Glasserman (1994), who uses techniques from Asmussen (1987).

4. Brief summary of related topics

(1) *Re-entrant flow shops.* How do we handle multi-product *re-entrant* flow lines, a very typical topology in semi-conductor fabrication facilities? (Re-entrant flow lines have attracted significant interest from the research community in recent years because of their direct applicability to semi-conductor fabrication.) This question was motivated by a fab visit at AMD and is studied by Bispo & Tayur (1996). Along the way we are able to study serial multi-product capacitated systems. The framework is similar to that described in § 2 above, but requires a far more detailed analysis.

(2) *Lead time quotation.* To compete effectively, suppliers are realizing that three aspects of lead times are important to customers – short, accurate and consistent – and that this 'service' has to be provided at minimum cost. Several improvements along the just-in-time (JIT) philosophy (and quality paradigms) have made production processes fairly reliable at many supplier plants. While negotiations with customers have reduced the variability of order streams, these variabilities continue to remain significant and further reductions seem unlikely given that the customers themselves face an intensely competitive marketplace with increasingly fickle and demanding (end) customers.

We concentrate on (in Kapuscinski & Tayur 1996b) the uncertainty of demand from multiple customer classes ($1 \leq i \leq M$) and consider production times to be deterministic. It is costlier to quote a longer lead time to a customer from a higher class than from a lower one; in particular, the cost is linear with rate m_i for class i customer. We impose the constraint that we will ship the order to the customer with a 100% reliability within the quoted lead time; we may ship earlier at no penalty, but we gain no benefit from the customer for this early shipment. Our original motivation to study this problem came from a company that produces 'rolls' that are used in steel mills. The goal was to find an easily implementable lead-time quoting rule, preferably graphical, that the sales force could use.

Briefly, we consider a finite-horizon, discrete-time production-inventory model with a single stage, single product that faces a stochastic demand from many customer classes in any period. Processing time is deterministic. In each period, after the demands are realized:

(a) we quote lead times to these demands,

(b) we make production decisions for this period, and

(c) we ship some material (sometimes earlier than its due date).

Items (2) and (3) above determine the inventory level at the end of the period, on which we pay a holding cost of h per unit (per period). Note that quoting a large lead time for a low margin customer (and so having a lot of capacity in anticipation of future higher margin customers) causes a penalty from this low margin customer, while a short quoted lead time would force a higher margin customer arriving in the near future to wait a little longer. This is the basic trade-off. We find a policy that is simultaneously optimal for 100% reliable quotation, production, inventory management and shipping. We show that for some cases the policy is easily implementable.

A simple model relating lead times, inventories and batching is studied by Karmarkar (1987); however, no quotation of due dates is considered there.

(3) *Co-operation in supply chains.* Industrial supplier–customer relations have undergone radical changes in recent years as the philosophy behind managing manufacturing systems is influenced by several Japanese manufacturing practices. As more organizations realize that successful in-house implementation of just-in-time alone will have limited effect, they are seeking other members of their supply chain to change their operations. This has resulted in a certain level of co-operation, mainly in the areas of *supply contracts* and *information sharing*, that was lacking before. This is especially true when dealing with customized products, and is most commonly seen between suppliers and their larger customers.

We incorporate (in Gavirneni *et al* 1996) information flow between a supplier (or producer) and a customer in a capacitated setting of a simple supply chain. The customer faces i.i.d. end-product demand, and the supplier has a finite capacity in each period. We consider three situations: (1) a traditional model where there is no information to the supplier prior to a demand to him except from past data; (2) the supplier has the information of the (s, S) policy used by the customer as well as the end-product demand distribution; and (3) full information about the state of the customer. Each of these leads to different

n-stationary demand processes as seen by the supplier. Study of these three models ables us to understand the relationships between capacity, inventory and information at e supplier level. We show that order up-to policies continue to be optimal for models th information flow for the finite horizon, the infinite horizon discounted and the infinite rizon average cost cases. We develop a quick recursive solution procedure to compute e optimal parameters when capacity is infinite. For the finite capacity case, we develop d validate Infinitesimal Perturbation Analysis (IPA), as well as show how the solution the uncapacitated system can be easily modified to obtain approximate values. Using ese solution procedures we estimate the savings at the supplier due to information flow d study when information is most beneficial by varying capacity, holding costs, demand stributions and $S - s$ values.

Summary

veral significant advances have occurred in the study of capacitated systems since 1991. is paper informally provides an introduction to topics, and some techniques. The reference list below is not exhaustive at all, but should provide a good starting point. At ist two very interesting developments (within a single product setting) are not considered in this informal review: (1) supply contracts and (2) international supply chains. See heller-Wolf & Tayur (1997), for example.

eferences

mussen S 1987 *Applied probability and queues* (New York: Wiley)

rtsekas D P 1987 *Dynamic programming: Deterministic and stochastic models* (Englewood Cliffs, NJ: Prentice-Hall)

spo C, Tayur S 1996 Re-entrant flow lines. GSIA Working Paper

izacott J, Shanthikumar J G 1993 *Stochastic models of manufacturing systems* (New York: Prentice-Hall)

ark A, Scarf H 1960 Optimal policies for a multi-echelon inventory problem. *Manage. Sci.* 6: 474–490

arallo F, Akella R, Morton T E 1994 A periodic review, production-planning model with uncertain capacity. *Manage. Sci.* 40: 320–332

hen M A, Lee H L 1990 Scale economics, manufacturing complexity, and transportation costs on supply chain facility networks. *J. Manuf. Oper. Manage.* 3: 269–292

hen M A, Moon S 1990 Impact of production: Scale economics, manufacturing complexity, and transportation costs on supply chain facility networks. *J. Manuf. Oper. Manage.* 3: 269–292

ebodt M, Graves S C 1985 Continuous review policies for a multi-echelon inventory problem with stochastic demand. *Manage. Sci.* 31: 1286–1299

dergruen A, Zipkin P 1984a Approximations of dynamic multilocation production and inventory problems. *Manage. Sci.* 30: 69–84

dergruen A, Zipkin P 1984b Computational issues in an infinite horizon multi-echelon inventory model. *Oper. Res.* 32: 218–236

dergruen A, Zipkin P 1986a An inventory model with limited production capacity and uncertain demands I: The average cost criterion. *Math. Oper. Res.* 11: 193–207

demands II: The discounted-cost criterion. *Math. Oper. Res.* 11: 208–215

Federgruen A, Zipkin P 1986 An inventory model with limited production capacity and uncertain

Federgruen A, Schweitzer P, Tijms H 1983 Denumerable undiscounted decision processes with unbounded rewards. *Math. Oper. Res.* 8: 298–314

Gavirneni S, Kapuscinski R, Tayur S 1996 Value of information in capacitated supply chains. GSIA Working Paper, CMU, Pittsburgh, PA 15213

Glasserman P 1991 *Gradient estimation via perturbation analysis* (Norwell, MA: Kluwer)

Glasserman P 1994 Bounds and asymptotics for planning critical safety stocks. Columbia University Working Paper (to appear in *Oper. Res.*)

Glasserman P, Tayur S 1994 The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* 42: 913–925

Glasserman P, Tayur S 1995 Sensitivity analysis for base stock levels in multi-echelon production-inventory system. *Manage. Sci.* 41: 263–281

Glasserman P, Tayur S 1996 A simple approximation for a multistage capacitated production-inventory system. *Naval Res. Log.* 43: 41–58

Graves S C, Rinnooy Kan A, Zipkin P (eds) 1992 *Logistics of production and inventory: Handbooks in operations research and management science* (Amsterdam: North Holland, Elsevier) vol. 4

Kamesam P V, Tayur S 1993 Algorithms for multistage, capacitated assembly systems with stochastic demand. IBM Report

Karlin S 1960a Dynamic inventory policy with varying stochastic demands. *Manage. Sci.* 6: 231–258

Karlin S 1960b Optimal policy for dynamic inventory process with stochastic demands subject to seasonal variations. *J. Soc. Indust. Appl. Math.* 8: 611–629

Kapuscinski R, Tayur S 1996a A capacitated production-inventory model with periodic demand. GSIA Working Paper (to appear in *Oper. Res.*)

Kapuscinski R, Tayur S 1996b 100% Reliable quoted lead times. GSIA Working Paper

Karmarkar U 1987 Lot sizes, lead times and in-process inventories, *Manage. Sci.* 33: 409–418

Langenhoff L J G, Zijm W H M 1989 An analytical theory of multi-echelon production/distribution systems. Working Paper, Eindhoven University of Technology, Department of Mathematics and Computer Sciences (to appear in *Stat. Neerlandica*)

Lee H L, Billington C, Carter B 1991 Gaining control of inventory and service through design for localization. Working Paper

Morton T 1978 The non-stationary infinite horizon inventory problem. *Manage. Sci.* 24: 1474–1482

Morton T, Pentico D 1995 The finite horizon non-stationary stochastic inventory problem. *Manage. Sci.* 41: 334–343

Muckstadt J, Lambrecht M, Luyten R 1984 Protective stocks in multistage production systems. *Int. J. Prod. Res.* 6: 1001–1025

Prabhu U 1965 *Queues and inventories* (New York: Wiley)

Prabhu U 1980 *Stochastic storage processes* (New York: Springer-Verlag)

Rosling K 1989 Optimal inventory policies for assembly systems under random demands. *Oper. Res.* 37: 565–579

Scheller-Wolf A, Tayur S 1997 Reducing international risk through quantity contracts. GSIA Working Paper, CMU, Pittsburgh

Schmidt C, Nahmias S 1985 Optimal policy for a two-stage assembly system under random demand. *Oper. Res.* 33: 1130–1145

eck C J, Van der Waal J 1991 The capacitated multi-echelon inventory system with serial structure: The average cost criterion. COSOR 91–39, Dept. of Math. and Comp. Sci., Eindhoven Institute of Technology, Eindhoven, The Netherlands

yur S 1993 Computing the optimal policy in capacitated inventory models. *Stochast. Models* 9:

n Houtum G J, Zijm W H M 1990 Computational procedures for stochastic multi-echelon production systems, Centre for Quantitative Methods, Nederlandse Philips Bedrijven, B.V.

n Houtum G J, Zijm W H M 1994 On multi-stage production/inventory systems under stochastic demand. *Int. J. Prod. Econ.* 35: 391–400

n Houtum G J, Inderfuth K, Zijm W H M 1995 Materials co-ordination in stochastic multi-echelon systems, University of Twente, Enschede

>kin P 1989 Critical number policies for inventory models with periodic data. *Manage. Sci.* 35: 71–80

On the optimality of exhaustive service policies in multiclass queueing systems with modulated arrivals and switchovers

Y NARAHARI* and N HEMACHANDRA

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India
e-mail: [hari,hema]@csa.iisc.ernet.in

Abstract. Consider a single-server multiclass queueing system with K classes where the individual queues are fed by K-correlated interrupted Poisson streams generated in the states of a K-state stationary modulating Markov chain. The service times for all the classes are drawn independently from the same distribution. There is a setup time (and/or a setup cost) incurred whenever the server switches from one queue to another. It is required to minimize the sum of discounted inventory and setup costs over an infinite horizon. We provide sufficient conditions under which exhaustive service policies are optimal. We then present some simulation results for a two-class queueing system to show that exhaustive, threshold policies outperform non-exhaustive policies.

Keywords. Multiclass-queue; MMPP arrivals; setup times; exhaustive policies.

1. Introduction

Hofri & Ross (1987) investigated the following scheduling problem in multiclass queues with setup times (see figure 1): There is a single server attending to two classes of customers from two queues fed by independent homogeneous Poisson processes. The service times for the two classes are drawn independently from the same (general) distribution. A switchover time with or without a constant monetary cost per switch made is involved whenever the server switches from one queue to the other. An inventory holding cost, linear in queue length and having the same rate at the two queues is also included. Two separate cost structures are considered:

(1) Sum of discounted switchover cost and inventory holding cost over an infinite horizon

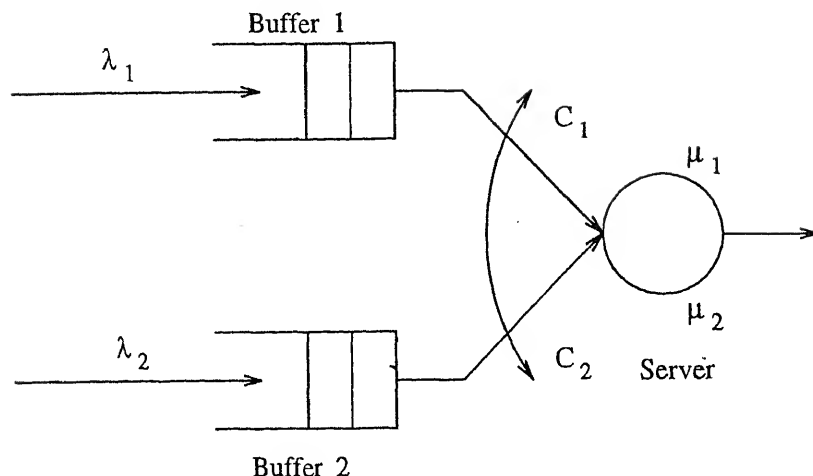(2) The long-run average switchover cost plus the inventory costs.

Figure 1.   A two-class queue with independent Poisson arrivals.

In both cases above, Hofri & Ross (1987) showed that the policies to minimize the cost:

(1) are necessarily *exhaustive*, i.e., server may switch to the other queue only when the current one is empty, and

(2) are likely to be *threshold* policies, i.e., the server switches (from an empty queue) only when the other reaches a critical size.

The first result, in fact, holds for three or more queues. Also, for the two-queue case, a detailed queueing-based procedure was given for computing the optimal thresholds.

   In this paper, we look at the following variant of the problem: The input streams to the two queues are no longer independent but are correlated in a Markov-modulated Poisson process (Fischer & Meier-Hellstern 1993) sense. (A Markov-modulated Poisson process can be informally described as an arrival process in which the Poisson arrivals have their rate modulated by a finite state, irreducible continuous-time Markov chain. In a K-state MMPP, the arrivals are Poisson at a certain rate (say, λ_i) so long as the CTMC is in state $i, \forall i \in \{1, \ldots, K\}$. This is a popular model for nonrenewal input to queues.) More specifically, we look into the system (depicted in figure 2) where the input streams are the two interrupted Poisson processes generated in the states of a two-state homogeneous, stationary modulating Markov chain. Consequently, the inputs to the queues are correlated. The motivation for considering input processes of this type comes from several situations in manufacturing and communication networks (Fischer & Meier-Hellstern 1993; Frost & Melamed 1994). For example, under Markovian switching, the output from a multiclass flexible machine constitutes an MMPP (Hemachandra & Narahari 1995). Similarly, the output of a failure-prone Markovian queue has the so-called on–off arrival feature – a Poisson output for an exponential amount of time when the machine is working and zero output for another exponential amount of time when it is not-working. We remark that the input to each class is an interrupted Poisson process which is stochastically equivalent to a hyperexponential renewal process (Fischer & Meier-Hellstern 1993).

   The problem and the investigations in this work are directly inspired by the work of Hofri & Ross (1987). Indeed, we follow the same line of argument and technical conditions
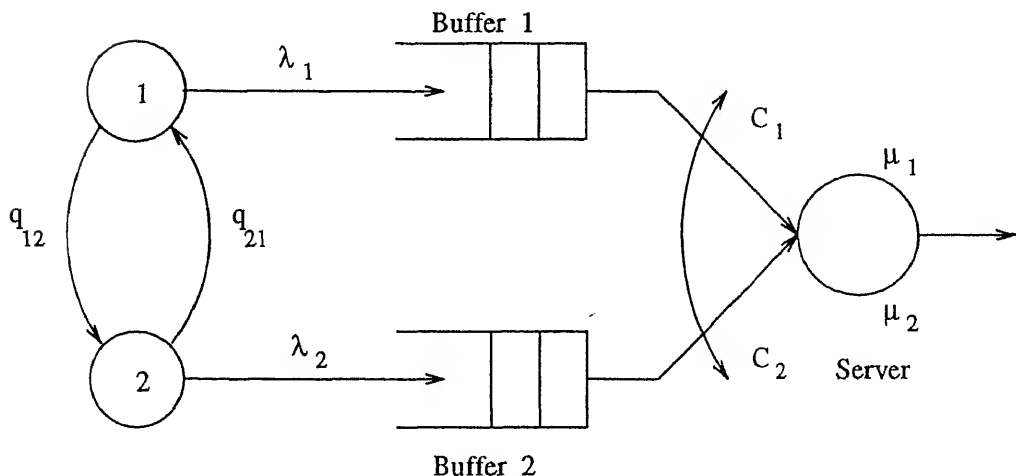
Figure 2. A two-class queue with modulated arrivals.

to corroborate our results. In this paper, we focus on the discounted cost criterion and investigate whether the optimal scheduling policies have to be necessarily exhaustive, like in the case of the Hofri–Ross formulation. Investigation of the correctness of this conjecture leads us to a set of sufficient conditions under which the conjecture is indeed true.

Apart from the seminal work of Hofri & Ross (1987), the literature that is relevant to this work is concerned with optimal scheduling in the presence of setups. The classic economic lot scheduling problem (Elmaghraby 1978) is the deterministic version of the multiclass scheduling problem with setups and has very elegant closed-form solutions. In the stochastic version which is more relevant here, two separate formulations exist. In the setup cost problem, a setup cost is incurred on each setup and in the setup time problem, a random setup time is incurred when the server switches class. The setup time problem is more realistic than the setup cost problem (Reiman & Wein 1994) but is also more difficult to solve.

Several researchers have studied the setup problem with more than two classes of customers, using a variety of techniques such as dynamic programming and heavy traffic queueing theory. A comprehensive survey appears in Reiman & Wein (1994) and Ravikumar (1996). All the authors however assume independent arrival processes into the individual queues. In § 2, we present the notation for the multiclass queueing model considered in this paper and describe the objective function to be optimized. In § 3, we investigate the conjecture that exhaustive policies are optimal and this leads to a set of sufficient conditions under which this conjecture is indeed true. In order to explore the validity of the conjecture, we carry out simulations on a two-class queueing system in § 4, which show that exhaustive, threshold policies outperform non-exhaustive policies.

2. The model

We consider a 2-class queue, representative of multiclass queues as a whole. Let the rates of the interrupted arrival processes to the two queues be λ_1 and λ_2 respectively, so that

the overall arrival process is a two-state MMPP. Let α_1 (α_2) be the rate from State 1 to State 2 (from State 2 to State 1) of the two-state Markov chain modulating the arrivals. Let π_1 and π_2 be the stationary probabilities of the modulating Markov chain so that $\pi_1 = \alpha_2/(\alpha_1 + \alpha_2)$ and $\pi_2 = \alpha_1/(\alpha_1 + \alpha_2)$. Service time distribution for a customer from either of the queues is the same and let $S_1, S_2, S_3, \cdots$ be the sequence of service times (*i.i.d.*) for customers of Queue 1 and $T_1, T_2, T_3, \cdots$ the sequence of service times (*i.i.d.*) for customers of Queue 2. Note that $\{S_n\}$ and $\{T_n\}$ are from the same service time distribution, with finite mean. Let $\{C_{1,n}\}$ and $\{C_{2,n}\}$ be two (possibly different) sequences of *i.i.d.* positive random variables with some general distributions, corresponding to the sequence of setup times at Queue 1 and Queue 2 respectively. The state of the system is then a four-tuple (u, x_1, x_2, v), where,

- u gives the status of the server taking values in the set $\{I_1, I_2, 1, 2, 12, 21\}$ where I_i means server is idle but setup for class i ($i = 1, 2$); 1 denotes that the server is currently serving Class 1 customers; 12 means that the server is switching from Class 1 to 2, and the other symbols have similar meaning.

- x_i is the number in the queue (including the one in service, if any) i ($i = 1, 2$), taking integer values.

- v takes values in the set $\{1, 2\}$ giving the state of the modulating Markov chain.

Let S be this state space. As in Hofri & Ross (1987), the decision epochs will be at service completions, switch completions and arrival instants. The actions at each such decision epoch are from the set $\mathcal{A}$ with $\mathcal{A} = \{S, C, I\}$ where these symbols successively mean that the server chooses to serve, change and idle. So the problem is a semi-Markov decision problem with state space S and action space $\mathcal{A}$ given above.

Let $X_i^{\pi, z}(t)$ be the number of customers in queue i ($i = 1, 2$) at time t, given that the initial state of the system is z, and a policy π is used. Let

$$X^{\pi, z}(t) = X_1^{\pi, z}(t) + X_2^{\pi, z}(t)$$

be the number in the system at time t. Let $R^{\pi, z}(t)$ be the number of switches made by the server up to time t under policy π. Then, the expected discounted cost, with discount factor $\beta > 0$, $V_\pi(z)$, incurred when a policy π is followed for a system starting in state z, is given by

$$V_\pi(z) = \mathbb{E}\left[\int_0^\infty e^{-\beta t} X^{\pi, z}(t) dt + a \int_0^\infty e^{-\beta t} dR^{\pi, z}(t)\right] \tag{1}$$

assuming that the holding costs per customer per unit time in both the queues are the same and are normalized to 1 and that a is the one-time switch charge levied at the start of each switch. A policy π' is said to be optimal if it attains

$$V(z) := V_{\pi'}(z) = \inf_\pi V_\pi(z) \quad \forall z \in S.$$

To suit the analysis ahead, we cast the objective function as in Harrison (1975): Let $A_i^{\pi, z}(t)$ be the number of arrivals up to time t and $D_i^{\pi, z}(t)$ the number of departures up to time t for the queue i ($i = 1, 2$) when a policy π is followed in a system starting at state z. Let x_1^0 and x_2^0 be the number in Queue 1 and Queue 2 respectively at time $t = 0$. Then, we have,

$$X_i^{\pi,z}(t) = x_i^0 + A_i^{\pi,z}(t) - D_i^{\pi,z}(t) \quad i = 1, 2, \tag{2}$$

so that from (1) and (2) we have,

$$V_\pi(z) = \frac{x_1^0 + x_2^0}{\beta} + \frac{\pi_1\lambda_1 + \pi_2\lambda_2}{\beta^2}$$
$$- \mathbb{E}\left[\int_0^\infty e^{-\beta t}(D_1^{\pi,z}(t) + D_2^{\pi,z}(t))\mathrm{d}t - a\tilde{R}^{\pi,z}(\beta)\right], \tag{3}$$

where $\tilde{R}^{\pi,z}(\beta)$ is the Laplace–Stieltjes transform (LST) of the distribution $R^{\pi,z}(t)$. Since,

$$\mathbb{E}\left[\int_0^\infty e^{-\beta t}(D_1^{\pi,z}(t) + D_2^{\pi,z}(t))\mathrm{d}t - a\tilde{R}^{\pi,z}(\beta)\right]$$
$$= \frac{1}{\beta}\mathbb{E}\left[\sum_{k\geq 1} e^{-\beta T_k^{\pi,z}} - a\beta\tilde{R}^{\pi,z}(\beta)\right],$$

where, $T_k^{\pi,z}$ is the kth departure epoch from the system under policy π, we have,

$$V_\pi(z) = \frac{x_1^0 + x_2^0}{\beta} + \frac{\pi_1\lambda_1 + \pi_2\lambda_2}{\beta^2} - \frac{1}{\beta}J_\pi(z),$$

with $J_\pi(z)$ defined as,

$$J_\pi(z) := \mathbb{E}\left[\sum_{k\geq 1} e^{-\beta T_k^{\pi,z}} - a\beta\tilde{R}^{\pi,z}(\beta)\right].$$

The optimizing problem is now equivalent to maximizing $J_\pi(z)$. If $J(z)$ is the supremum of $J_\pi(z)$ over all policies, then by Lippman (1973), there is an optimal pure policy which achieves $J(z)$.

3. Optimality of exhaustive policies

Let g be any non-exhaustive policy; we shall argue below that the performance of any such policy could be improved upon. This means that non-exhaustive policies are suboptimal. Thus, optimal policies have to be necessarily exhaustive.

First, we synthesize from the existing arrival process, a new arrival process that depends on the scheduling policy followed. Let f be any such policy.

Let $N_1(t)$ and $N_2(t)$ be the arrival processes to Queues 1 and 2. Assume that $\bar{N} := (\bar{N}_1(t), \bar{N}_2(t))$ is stochastically equivalent but independent of $N(t) := (N_1(t), N_2(t))$. Let $(\Omega, \mathcal{F}, P)$ support the independent processes $N(t)$, $\bar{N}(t)$ and the sequences $\{S_n\}$, $\{T_n\}$, $\{C_{1,n}\}$, $\{C_{2,n}\}$. As in Hofri & Ross (1987), form a new arrival process from these $N(t)$ and $\bar{N}(t)$ by collecting customers from $N(t)$ or $\bar{N}(t)$ as described below. Let τ^f be the first instant that the server begins to serve a customer from Queue 1 under policy f. Let S_1 be the service time of this customer from Queue 1. Define a new arrival process $A^f(t)$ as:

$$A^f(t) = N(t), \quad 0 \leq t \leq \tau^f,$$
$$= N(\tau^f) + \bar{N}(t - \tau^f), \quad \tau^f \leq t \leq \tau^f + S_1, \tag{4}$$
$$= \bar{N}(S_1) + N(t - S_1), \quad t \geq \tau^f + S_1.$$

That is, collect customers from $N(t)$ until τ^f, then on from $\bar{N}(t)$ until $\tau^f + S_1$; and revert back to $N(t)$ indefinitely after $\tau^f + S_1$. Note that $A^f(t)$ is defined pathwise and its construction depends on policy f.

Since g is non-exhaustive, there exists some state, z, such that either g causes the server to idle at a nonempty queue or abandons a nonempty queue. Without loss of generality, assume that this occurs at Queue 1. Then, if $g(u, x_1, x_2, v)$ denotes the scheduling decision in the state (u, x_1, x_2, v),

$$\exists x_1 > 0 \quad \ni g(1, x_1, x_2, v) \neq S,$$

i.e.,

$$g(1, x_1, x_2, v) \in \{C, I\} \quad \text{with} \quad x_1 > 0.$$

Let this $z_0 := (1, x_1, x_2, v)$ be the initial state in the sequel. Let $Z^{f, z_0}(t)$ denote the state of the system at time t when a policy f is followed by the system starting in a state z_0. Write $Z^\pi(t)$ for $Z^{\pi, z_0}(t)$.

We now exhibit a policy π that outperforms the policy g in minimizing the total cost. Policy π says that the system beginning in state z_0 should first *immediately* serve a job from Queue 1 (let S_1 be the service time of this customer). Then, π follows policy g for a random amount of time τ^g from S_1 to $\tau^g + S_1$ with a lag of S_1. From $\tau_g + S_1$ onwards it simply follows policy g with no delay. So, $\tau^\pi = 0$. Writing out $A^\pi(t)$ and $A^g(t)$, the arrival processes corresponding to policies π and g using (4), we have that they are related as:

$$\begin{aligned}
A^\pi(t) &= A^g(t + \tau^g) - A^g(\tau^g), \quad 0 \leq t \leq S_1, \\
&= A^g(t - S_1) + A^\pi(S_1), \quad S_1 \leq t \leq \tau^g + S_1, \\
&= A^g(t), \quad t \geq \tau^g + S_1.
\end{aligned}$$

Note that the departure processes, $D^\pi(t)$ and $D^\pi(t)$, are related as

$$\begin{aligned}
D^\pi(t) &= 0, \quad 0 \leq t \leq S_1, \\
&= 1 + D^g(t - S_1), \quad S_1 \leq t \leq \tau^g + S_1, \\
&= D^g(t), \quad t \geq \tau^g + S_1.
\end{aligned} \tag{5}$$

Essentially, we have states coupled (stochastically) by the policies g and π in such a way that

$$Z^\pi(t) = Z^g(t) \quad \text{for } t > \tau^g + S_1.$$

Here is the crucial assumption.

Assumption A.   Let the arrival processes $A^\pi(t)$ and $A^g(t)$ formed above be stochastically equivalent, so that the systems when driven by any of these processes have the same performance measures when the same policy is used for both the systems.

Remark.   This assumption is satisfied by independent Poisson processes as noted by Hofri

Let (i) at least one of the values a, $\mathbb{E}[C_1]$, $\mathbb{E}[C_2]$ be strictly greater than zero and (ii) the processes $A^\pi(t)$ and $A^g(t)$ be equal in law. Then, any non-exhaustive policy g is non-optimal.

Proof. Using (5) above, we show that the difference,

$$
J_\pi(z_0) - J_g(z_0)
$$
$$
= \beta \mathbb{E}\left[\int_{t \geq 0} e^{-\beta t}[D^\pi(t) - D^g(t)]dt\right] - a\beta \mathbb{E}\left[\int_{t \geq 0} e^{-\beta t}[dR^\pi(t) - dR^g(t)]\right],
$$
$$
\tag{6}
$$

is positive. This difference has two terms on the RHS, the first one capturing the holding cost and the other the switching cost.

Let σ^g be the first time that g switches from Queue 1 when the initial state is z_0. On each of the sets

$$
\Lambda_1 := \{\tau^g < \sigma^g\}, \qquad \Lambda_2 := \{\tau^g = \sigma^g = \infty\}, \qquad \Lambda_3 := \{\tau^g \geq \sigma^g < \infty\},
$$

we compute this difference assuming that $\mathbb{E}[C_2] > 0$ or $a > 0$.

Case 1 $(\tau^g < \sigma^g)$. Here we have $R^\pi(t) = R^g(t)$, $t \geq 0$ and hence the difference in switching costs as captured by the second term above is zero. Since

$$
D^\pi(t) = D^g(t) + 1, \quad t \in [S_1, \tau^g + S_1],
$$
$$
= D^g(t), \quad t \notin [S_1, \tau^g + S_1],
$$

we have,

$$
\beta \mathbb{E}\left[\int_{t \geq 0} e^{-\beta t}[D^\pi(t) - D^g(t)]dt | \Lambda_1\right] = \mathbb{E}[e^{-\beta S_1}(1 - e^{-\beta \tau^g}) | \Lambda_1]
$$
$$
= \mathbb{E}[\mathbb{E}[e^{-\beta S_1}(1 - e^{-\beta \tau^g}) | \Lambda_1, \tau^g] | \Lambda_1]
$$
$$
= \mathbb{E}[(1 - e^{-\beta \tau^g})\mathbb{E}[e^{-\beta S_1} | \Lambda_1, \tau^g] | \Lambda_1]
$$
$$
= \mathbb{E}[e^{-\beta S_1}]\mathbb{E}[(1 - e^{-\beta \tau^g}) | \Lambda_1]
$$

Now, $\mathbb{E}[e^{-\beta S_1}] =: \tilde{S}(\beta)$ and since τ^g is strictly positive a.s. we have this term greater than zero and thus (6) is positive in Case 1.

Case 2 $(\tau^g = \sigma^g = \infty)$. Policy g simply idles forever and hence the difference in switching costs is zero and the difference in holding costs is given by

$$
\beta \mathbb{E}\left[\int_0^\infty e^{-\beta t}[D^\pi(t) - D^g(t)]dt | \Lambda_2\right] = \mathbb{E}[e^{-\beta S_1} | \Lambda_2] = \tilde{S}(\beta)
$$

so that, here also (6) is positive.

Case 3 ($\tau^g \geq \sigma^g < \infty$). Many scenarios are possible here with policy g deciding to opt for a switching before taking up a job from Queue 1 (in finite time a.s.). In all of them, however, $R^\pi(t)$ lags behind $R^g(t)$ by S_1 till $\tau^g + S_1$ and from then on they are the same. Here, we show that the difference (6) is strictly positive by showing that the first term (holding cost difference) is strictly positive if $\mathbb{E}[C_2] > 0$, and the negative of the second term (switching cost difference) is also strictly positive, if $a > 0$.

First, it follows as in Hofri & Ross (1987) again, that the negative of the switching cost difference in (6) above can be lower bounded,

$$-a\beta\mathbb{E}\left[\int_0^\infty e^{-\beta t}[dR^\pi(t) - dR^g(t)]|\Lambda_3\right] \geq a\beta\mathbb{E}[e^{-\beta\sigma^g}(1 - e^{-\beta S_1})|\Lambda_3]$$

and hence, is strictly positive if $a > 0$.

If $\mathbb{E}[C_2] > 0$, it follows that the holding cost difference in (6) is strictly positive as shown below. Write

$$\beta\mathbb{E}\left[\int_0^\infty e^{-\beta t}[D^\pi(t) - D^g(t)]dt|\Lambda_3\right]$$

$$= \beta\int_0^\infty e^{-\beta t}\left(\sum_{k\geq 1}[\mathbb{P}(D^\pi(t) \geq k, \Psi|\Lambda_3) - \mathbb{P}(D^g(t) \geq k, \Psi|\Lambda_3)]\right)dt$$

where $\Psi := \{t \leq S_1 + \tau^g\}$. Note that departure processes are stochastically equal for $t > S_1 + \tau^g$. With T_i^g denoting the ith departure epoch from the system under policy g, ($T_0^g := 0$), set $U_i := T_i^g - T_{i-1}^g$. Then, with $U_k \geq_{st} S_1$, we have,

$$\mathbb{P}(U_1 + U_2 + \cdots + U_{k-1} + S_1 \leq t) \geq \mathbb{P}(U_1 + U_2 + \cdots + U_k \leq t)$$
$$\forall t \geq 0, \ k \geq 0$$
$$\Rightarrow \mathbb{P}(D^\pi(t) \geq k, \Psi|\Lambda_3) \geq \mathbb{P}(D^g(t) \geq k, \Psi|\Lambda_3) \ \forall t \geq 0, \ k \geq 0$$

so that, the above is indeed positive. If, $\mathbb{E}[C_2] > 0$, then from the fact that U_1 includes the service time of a job as well as switchtime or idling time, we have,

$$\mathbb{P}(S_1 \leq t, \Psi|\Lambda_3) \geq \mathbb{P}(U_1 \leq t, \Psi|\Lambda_3) + \epsilon,$$

for some $\epsilon > 0$ and t in an open interval. This means that

$$\mathbb{P}(D^\pi(t) \geq 1, \Psi|\Lambda_3) \geq \mathbb{P}(D^g(t) \geq 1, \Psi|\Lambda_3) + \epsilon,$$

for some $\epsilon > 0$ and t in an open interval. So, finally, we have (6) strictly positive under the hypothesis.

Remark 1. It is easy to see that the above arguments go through for queues with more than two classes; we require that at least one of a, $\{\mathbb{E}[C_{i,j}]\}$ be strictly positive.

Remark 2. The proof above uses the stochastic coupling idea and the requirement on the part of the arrival processes, as set out in assumption A, facilitates coupling of the evolution of the systems following the policies *pathwise*, after the coupling epoch, $\tau^g + S_1$. Since the technical nature of MMPP is not used in the above proof, we conjecture that the above result is valid for a much more general arrival process that satisfies assumption A.

Remark 3. Policy π is *more* exhaustive than policy g in the sense that it serves one job from the *same* queue in the very beginning. As the above argument is sample pathwise, by induction, this means that exhaustive policies are better than non-exhaustive policies.

Remark 4. We left open here the decisions the server has to take when the queue to which it is attached becomes empty. One class of policies to use in this context is that of the *threshold policies* in which Hofri & Ross (1987) show that such policies are optimal when arrivals are Poisson.

Remark 5. Verifying the sufficiency condition involving the arrival process (Assumption A) is difficult; indeed it may be necessary to pursue another track of investigation to get around this problem (M Hofri 1996, personal communication).

4. Simulation results

Since it is difficult to verify the validity of the sufficient conditions concerning the equivalence of the arrival processes (assumption A), we simulated quite a few systems operating under different policies and looked at the cost incurred by them. This way, we investigated

Table 1. Setup costs and inventory holding costs for a bursty scenario.

$n_1\,n_2$	$N_1\,N_2$	SW_1	SW_2	Switching cost	Holding cost	Total cost
1 1	**2 3**	**1337.6**	**1252.4**	**2590.0**	**798.0**	**3388.0**
1 2	2 3	1342.5	1267.4	2609.9	1526.3	4136.2
1 3	2 3	1456.1	1404.4	2860.5	2228.2	5088.7
2 1	2 3	1375.8	1506.3	2882.1	1656.1	4538.2
3 1	2 3	1817.8	1909.5	3727.3	2142.1	5869.4
2 2	2 3	1387.9	1532.7	2920.6	2329.9	5250.5
2 3	2 3	1504.2	1675.4	3179.6	3008.8	6188.2
1 1	**2 2**	**1342.5**	**1275.4**	**2617.9**	**591.1**	**3209.0**
1 2	2 2	1456.1	1412.3	2868.4	1325.2	4193.6
1 3	2 2	1785.4	1755.5	3540.9	2190.7	5731.7
2 1	2 2	1387.4	1540.7	2928.6	1394.6	4323.2
3 1	2 2	2001.6	2132.1	4133.7	1780.2	5913.9
1 1	**3 2**	**1160.9**	**1101.5**	**2262.4**	**732.5**	**2994.9**
1 2	3 2	1259.0	1224.1	2483.1	1507.8	3991.0
1 3	3 2	1401.4	1377.6	2779.0	2441.1	5220.1
1 5	3 2	1401.9	1379.2	2781.1	4154.4	6935.5
2 1	3 2	1112.3	1275.4	2387.7	1511.3	3899.0
2 2	3 2	1225.9	1412.3	2638.3	2245.4	4883.7
2 3	3 2	1555.2	1755.5	3310.7	3111.0	6421.7
2 4	3 2	1633.7	1835.9	3469.6	3972.3	7441.9
3 1	3 2	1202.3	1357.6	2559.9	2116.1	4676.0
3 2	3 2	1318.6	1499.7	2818.3	2831.4	5649.8
3 3	3 2	1869.5	2065.4	3934.9	3628.2	7563.1
4 1	3 2	1816.0	1949.0	3765.0	2501.7	6266.7
4 2	3 2	2187.1	2348.6	4535.7	3073.3	7608.9
5 1	3 2	2237.9	2343.8	4580.8	2907.0	7487.8
5 2	3 2	3018.9	3158.6	6177.6	3428.1	9605.6

Table 2. Setup and inventory holding costs for exhaustive policies.

$n_1 n_2$	$N_1 N_2$	SW_1	SW_2	Switching cost	Holding cost	Total cost
1 1	1 1	1734.3	1683.7	3418.0	294.1	3712.1
1 1	1 2	1618.1	1540.7	3158.8	474.4	3633.2
1 1	1 3	1606.0	1506.3	3112.3	735.9	3848.2
1 1	1 4	1257.2	1149.2	2406.4	964.8	3371.1
1 1	1 5	1217.1	1096.6	2313.6	1256.0	3569.6
1 1	2 1	1456.1	1412.7	2868.8	433.9	3302.7
1 1	2 2	1342.5	1275.4	2617.8	591.1	3209.0
1 1	2 3	1337.6	1252.4	2589.9	797.4	3387.3
1 1	2 4	1232.9	1137.4	2370.4	936.8	3307.1
1 1	2 5	1184.9	1075.7	2260.6	1233.0	3493.7
1 1	3 1	1259.0	1224.5	2483.5	616.5	3100.0
1 1	**3 2**	**1160.9**	**1101.5**	**2262.4**	**732.5**	**2994.9**
1 1	3 3	1160.1	1082.4	2242.5	930.6	3173.1
1 1	3 4	1056.1	970.1	2026.2	1008.6	3034.8
1 1	3 5	1008.5	910.2	1918.7	1268.7	3187.4
1 1	4 1	1122.2	1091.3	2213.4	940.6	3154.1
1 1	4 2	1024.7	971.6	1996.3	1027.7	3024.0
1 1	4 3	1019.7	950.9	1970.5	1209.7	3180.1
1 1	4 4	1019.2	939.8	1959.0	1150.3	3109.3
1 1	4 5	1001.0	905.5	1906.5	1324.0	3230.5
1 1	5 1	1096.3	1067.1	2163.8	1035.9	3199.3
1 1	5 2	1016.4	965.3	1981.7	1094.2	3075.9
1 1	5 3	1015.8	949.0	1964.7	1263.5	3228.2
1 1	5 4	1005.1	927.3	1932.4	1201.0	3133.4
1 1	5 5	998.8	905.2	1904.0	1371.3	3275.3

the performance of the exhaustive policies. Also, we confine our experimentation to the class of exhaustive, threshold policies motivated by the results of Hofri & Ross (1987) concerning the optimality of threshold policies in the case of independent Poisson arrivals. We remark here that exhaustive, threshold policies may not be the optimal policies.

A typical policy we considered is captured by the four tuple (n_1, n_2, N_1, N_2). Here, n_1 and n_2 are *idle* thresholds while N_1 and N_2 are *switching* thresholds. n_1 (n_2) is the minimum number of jobs required in Queue 1 (Queue 2) for the server to process a job from that queue when attached to the queue. In other words, even though the server is attached to Queue 1 (Queue 2) and the number of customers in that queue is less than n_1 (n_2), it will not process any job from that queue. Note that a service policy is exhaustive if $n_1 = n_2 = 1$. N_1 (N_2) is the minimum number of jobs required in Queue 1 (Queue 2) for the server, currently attached to Queue 2 (Queue 1), to switch to Queue 1 (Queue 2). N_1 and N_2 are called the switching thresholds.

Note that, in view of the exhaustive nature of the policies, n takes precedence over N. So, in an empty system, the server when attached to a queue, say Queue 1, potentially idles there till the number in this queue is greater than or equal to n_1. If, during this idling period, the number in Queue 2 becomes at least N_2, it switches. Similarly, the server continues

le 3.   Setup and inventory holding costs for a non-bursty scenario.

$_2$	N_1 N_2	SW_1	SW_2	Switching cost	Holding cost	Total cost
	4 5	**692.3**	**550.8**	**1243.1**	**1189.5**	**2432.6**
	4 5	694.2	556.6	1250.9	1766.7	3017.6
	4 5	735.0	599.6	1334.5	2329.9	3664.4
	4 5	863.1	732.0	1595.2	2814.6	4409.4
	4 5	764.0	618.6	1382.6	2098.3	3480.9
	4 5	775.9	634.8	1410.6	2653.2	4063.8
	4 5	800.2	663.4	1463.7	3238.8	4702.4
	4 5	929.0	796.1	1725.1	3702.2	5427.4
	4 5	886.5	737.9	1624.4	3088.9	4713.3
	4 5	898.4	754.8	1653.2	3624.8	5278.1
	4 5	999.8	857.6	1857.5	4175.7	6033.1
	4 5	1129.3	992.0	2121.3	4606.1	6727.4
	4 5	888.3	737.9	1626.2	4027.6	5653.8
	4 5	900.2	754.9	1655.1	4563.0	6218.1
	4 5	1001.7	857.7	1859.4	5113.2	6972.5
	4 5	1131.7	992.0	2123.7	5537.7	7661.3
	4 5	1222.0	1049.4	2271.4	5140.6	7412.0
	4 5	1336.9	1167.2	2504.1	5705.8	8209.9
	4 5	1501.0	1342.8	2843.8	6118.0	8961.8
	4 5	1323.9	1146.8	2470.7	6124.7	8595.4
	4 5	1442.9	1270.0	2712.9	6666.4	9379.3
	4 5	1711.6	1542.7	3254.3	6968.2	10222.6
	4 4	**835.9**	**699.1**	**1535.1**	**949.2**	**2484.3**
	4 4	876.7	742.2	1618.9	1559.7	3178.5
	4 4	1004.8	875.6	1880.4	2074.3	3954.7
	4 4	1012.0	890.8	1902.9	2746.0	4648.9
·	4 4	917.7	777.3	1695.0	1834.3	3529.3
	5 4	**831.2**	**698.8**	**1530.0**	**1062.2**	**2592.2**
	5 4	871.5	741.8	1613.4	1685.7	3299.1
	5 4	999.2	875.3	1874.5	2217.1	4091.6
	5 4	1006.3	890.5	1896.8	2890.0	4786.8
	5 4	835.0	699.2	1534.1	1936.9	3470.9
	5 5	**688.8**	**550.4**	**1239.2**	**1268.1**	**2507.3**
	5 5	690.7	556.3	1245.0	1847.2	3094.2
	5 5	731.0	599.2	1330.2	2423.3	3753.5
	5 5	691.3	550.8	1242.1	2177.2	3419.4

Examples

re, we look at the performance costs incurred by two systems as they follow policies
ich are more and more exhaustive, i.e., with n_1, n_2 or both, becoming less. For the
tems we considered, the simulations show that the exhaustive policies incur the least
st.

ample 1. (Bursty arrivals). Recall that a central motivation for the use of MMPP as an
ival process is to capture bursty arrivals to a queue. We wish to look at such a scenario
w in the context of service and scheduling policies. Let $\alpha_1, \alpha_2, \lambda_1$, and λ_2 have the usual
nificance. The numerical data are: $\alpha_1 = 1/10, \alpha_2 = 1/60, \lambda_1 = 0.2, \lambda_2 = 0.04$. Let
 setup time be deterministic at a value 10 for both the queues and the mean service time
terministic at 0.50. The discount factor $\beta = 0.001$, while the setup cost $a = 250$.

Table 4. Setup and inventory holding costs for exhaustive policies.

n_1 n_2	N_1 N_2	SW_1	SW_2	Switching cost	Holding cost	Total cost
1 1	1 1	1486.9	1360.7	2847.5	183.0	3030.5
1 1	1 2	1426.8	1286.4	2713.2	334.7	3047.5
1 1	1 3	1149.1	1001.1	2150.9	644.3	2794.5
1 1	1 4	1047.6	897.4	1945.0	782.3	2727.3
1 1	1 5	892.4	737.9	1630.4	1085.1	2715.5
1 1	1 6	813.7	653.6	1467.3	1346.2	2813.5
1 1	2 1	1482.0	1360.5	2842.5	247.0	3089.5
1 1	2 2	1422.2	1286.3	2708.6	395.4	3103.9
1 1	2 3	1145.3	1001.1	2146.4	694.0	2840.4
1 1	2 4	1043.9	897.3	1941.2	831.3	2772.6
1 1	2 5	890.2	737.9	1628.0	1114.6	2742.7
1 1	2 6	811.9	653.5	1465.4	1367.2	2832.7
1 1	3 1	1236.1	1119.6	2355.7	394.3	2750.0
1 1	3 2	1218.9	1090.5	2309.4	483.1	2792.5
1 1	3 3	943.0	806.8	1749.8	744.0	2493.9
1 1	3 4	918.6	777.3	1695.9	846.6	2542.5
1 1	3 5	765.0	618.6	1383.6	1110.7	2494.3
1 1	3 6	681.7	533.0	1214.8	1338.9	2553.6
1 1	3 7	644.5	494.2	1138.7	1478.8	2617.6
1 1	4 1	1013.5	900.2	1913.7	543.8	2457.5
1 1	4 2	1006.3	882.2	1888.5	611.0	2499.5
1 1	4 3	876.7	743.0	1619.7	824.1	2443.8
1 1	4 4	835.9	699.1	1535.1	949.1	2484.2
1 1	**4 5**	**692.3**	**550.8**	**1243.2**	**1189.6**	**2432.8**
1 1	4 6	615.7	469.7	1085.5	1392.7	2478.2
1 1	4 7	579.1	430.9	1010.0	1522.9	2532.9
1 1	4 8	576.2	424.6	1000.7	1646.8	2647.6
1 1	4 9	499.4	348.2	.847.6	1968.9	2816.6
1 1	5 1	1007.5	899.8	1907.3	698.9	2606.1
1 1	5 2	1000.4	881.8	1882.2	764.7	2646.9
1 1	5 3	871.5	742.6	1614.1	950.0	2564.1
1 1	5 4	831.21	698.8	1530.0	1062.1	2592.1
1 1	5 5	688.76	550.4	1239.2	1268.0	2507.2
1 1	5 6	612.5	469.7	1082.1	1473.2	2555.4
1 1	6 1	986.0	881.7	1867.7	902.5	2770.2
1 1	6 2	979.0	865.0	1844.0	957.3	2801.3
1 1	6 3	850.3	726.3	1576.5	1131.5	2708.0
1 1	6 4	827.3	698.0	1525.4	1150.9	2676.3
1 1	6 5	685.31	550.0	1235.3	1352.4	2587.8

This system was simulated over a sufficiently long simulation run length of 10,000 time units and the total cost (as given by (1)) was computed for different sets of values of n_1, n_2, N_1, N_2. It was found that the run length of 10,000 was an adequate indicator of the infinite horizon discounted cost since the accumulated cost became virtually saturated for this run length. The simulation clock was made to start ticking only after ensuring that the modulating Markov chain was well into the steady state.

Table 1 shows for different values of n_1, n_2, N_1, N_2 the values of the accumulated discounted switching costs from Queue 1 and Queue 2, respectively, as SW_1, SW_2, the total switching cost $(SW_1 + SW_2)$, the accumulated (discounted) inventory holding cost, and the overall discounted cost. Observe that three sets of values have been explored for

and N_2, namely, $N_1 = 2$, $N_2 = 3$; $N_1 = 2$, $N_2 = 2$; $N_1 = 3$, $N_2 = 2$. From table 1
see that in all these three sets, exhaustive policies as given by $n_1 = n_2 = 1$ have the
t performance.

Next, having seen that exhaustive policies perform better, we investigate the effective-
s of threshold switching by looking at the cost incurred for different values of thresholds
and N_2 (see table 2). For this system, the threshold $N_1 = 3$, $N_2 = 2$ offers the least
rall discounted cost among threshold policies considered.

ample 2. (Non-bursty arrivals). This scenario comes up when the streams of jobs have
nparable intensities with a correlation as captured by the 2-state Markov chain. We
/e here, $\alpha_1 = \alpha_2 = 1/60$, $\lambda_1 = 0.20$, $\lambda_2 = 0.10$. The service is deterministic at value
0. The setup cost is 200 while the setup times are deterministic at values 5 and 5. The
count factor $\beta = 0.001$. The notation in tables 3 and 4 is as in the previous example.
le 3 demonstrates that exhaustive policies are better when thresholds are taken as (4,5);
4); (5,4); and (5,5). Then, fixing the policy as exhaustive, i.e., with $n_1 = n_2 = 1$, we
rch for the thresholds which offer the least overall discounted cost; Table 4 shows that
thresholds $N_1 = 4$, $N_2 = 5$ offer the best performance among the policies considered.

Conclusions

r a single-server multiclass queue fed by the streams of an MMPP, with non-preemptive
itchovers involving a switchover cost and/or a switchover time, we argued that ex-
ustive policies are optimal as long as the arrival process satisfies a technical condition
;arding its law and the service times of the various classes are the same. The verification
this technical requirement on the part of the arrival process appears to be difficult. Hofri
Ross (1987) showed that, when the queue is fed by independent Poisson streams, these
licies are optimal; here, this technical condition is true. To test our claim, we carried out
ite an extensive simulation of two examples in the case of a two-class queue. Exhaus-
e policies were found to outperform non-exhaustive policies in all cases. Further, these
nulations showed that threshold policies are the best performing among all exhaustive,
eshold policies. Apart from having a direct argument about the optimality of exhaustive
licies, the question of effectiveness of the threshold policies needs to be looked into in
ure.

:ferences

naghraby S E 1978 The economic lot scheduling problem (ELSP): Review and extentions.
Manage. Sci. 28: 587–598

Fischer W, Meier-Hellstern K 1993 The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 18: 149–171

Frost V S, Melamed B 1994 Traffic modeling for telecommunication networks. *IEEE Commun. Mag.* 10: 70–81

Harrison J M 1975 A priority queue with discounted linear costs. *Oper. Res.* 23: 260–269

Hemachandra N, Narahari Y 1995 A linear programming approach to Markovian switching among Poisson streams to a queueing system. Technical report, Department of Computer Science and Automation, Indian Institute of Science, Bangalore

Hofri M, Ross K W 1987 On the optimal control of two queues with server setup times and its analysis. *SIAM J. Comput.* 16: 399–420

Lippman S 1973 Semi-Markov decision processes with unbounded rewards. *Manage. Sci.* 1 717–731

Ravikumar K 1996 *Dynamic and stochastic scheduling of multi-product queues with setups A diffusion approach.* Ph D dissertation, Department of Computer Science and Automation Indian Institute of Science, Bangalore

Reiman M I, Wein L M 1994 Dynamic scheduling of a two-class queue with setups. Technical report, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA

An extension of modified-operational-due-date priority rule incorporating job waiting times and application to assembly job shop

P G AWATE* and P V SARAPH+

*Industrial Engineering & Operations Research, Indian Institute of Technology, Bombay 400 076, India
+Present address: Institute of Development Studies, University of Sussex, Brighton, BN1 9RE, UK
e-mail: [awatepg, prasadv]@me.iitb.ernet.in

Abstract. The well-known priority dispatching rule MOD (Modified Operational Due Date) in job shop scheduling considers job urgency through ODD (Operational Due Date) and also incorporates SPT(Shortest Processing Time)-effect in prioritising operationally late jobs; leading to robust behaviour in Mean Tardiness (MT) with respect to tightness/looseness of due dates. In the present paper, we study an extension of the MOD rule using job-waiting-time based discrimination among operationally late jobs to protect long jobs from excessive delays by incorporating an 'acceleration property' into the scheduling rule. Formally, we employ a weighted-SPT dispatching priority index of the form: (Processing time)/(Waiting time)$^\alpha$ for operationally late jobs, while the priority index is ODD for operationally non-late jobs; and the latter class of jobs has a lower priority than the former class.

In the context of Assembly Job Shop scheduling, some existing literature includes considerable focus around the concept of 'Staging Delay', i.e., waiting of components or sub-assemblies for their counterparts for assembly. Some existing approaches attempt dynamic anticipation of staging delay problems and re-prioritisation of operations along converging branches. In the present paper, rather than depending on such a centralised and largely backward scheduling approach, we consider a partially decentralised approach, endowing jobs with a priority index yielding an 'acceleration property' based on a 'look-back' in terms of waiting time, rather than 'look-ahead'. For the particular case, in our proposed rule, when α is set at zero and when all jobs at a machine are operationally late, our rule agrees with MOD as both exhibit the SPT effect.

In simulation tests of our priority scheme for assembly job shops, in comparison with leading heuristics in literature, we found our rule to be particu-

larly effective in: (1) minimising conditional mean tardiness, (2) minimising 99-percentile-point of the tardiness distribution, through proper choice of α. We also exploit an interesting duality between the scheduling and queueing control versions of the problem. Based on this, some exact and heuristic analysis is given to guide the choice of α, which is also supported by numerical evidence.

Keywords. Scheduling; assembly job shops; queueing; waiting times; dynamic priorities.

1. Introduction

In job shop scheduling, meeting due dates is a crucial objective. Due date related priority dispatching rules essentially are based on a primarily backward scheduling approach. This can, however, lead to some uneven loading of machines over time. This cannot be avoided even if due dates are internally set because in the setting of a dynamic job shop (that is to be governed by dispatching priority rules), no sequential due date setting procedure for jobs (in the order of their arrival) can anticipate all possible conflicts among jobs for resources over time, even in a non-probabilistic situation.

Uneven loading of machines over time can be associated with uneven rates of job flows between machines and increased waiting times for jobs, due to greater variability in the arrival process for jobs coming to a machine. The well-known Modified Operation Due Date (MOD) rule proposed and studied by Baker (1984) judiciously incorporates the SPT-effect when there exist operationally late jobs at a machine. This leads towards locally maximising job flow rates past the machine, and such mobility can partially help to lessen the high ups and downs in machine loading over time in a purely backward scheduling approach based on ODD alone. Thus, although SPT is a local dispatching rule (associated with minimising mean flow time past a machine), the SPT-effect as built into the MOD rule indeed turns out to give low Mean Tardiness (MT) in job shop scheduling as such. The robustness of MOD (in respect of varying tightness of due dates) relates to the fact that ODD relates to minimising maximal operational lateness, while for operationally late jobs, minimising their mean flow time correlates with minimising their mean tardiness. Baker (1984) had also anticipated consideration of MOD rule in assembly job shops.

In connection with our observations on the impact of SPT-effect as above in a job shop, it may also be stated that a study by Russell & Taylor (1985) has suggested that the performance of a pure SPT rule improves as the product structure (in assembly job shop) grows taller (where in fact there would exist a greater need for job mobility in the sense described above). Concern over the uneven machine loading due to purely backward scheduling strategies has also been expressed in the work of Wein & Chevalier (1992).

On the whole, the incorporation of the SPT effect as in the MOD rule due to Baker (1984) leads to beneficial effects in job shop scheduling with due dates, and this can be interpreted or motivated in different ways as indicated above.

In assembly job shops (as opposed to pure or component job shops) there is the phenomenon of staging delay (Adam *et al* 1987), i.e., waiting of a component for the
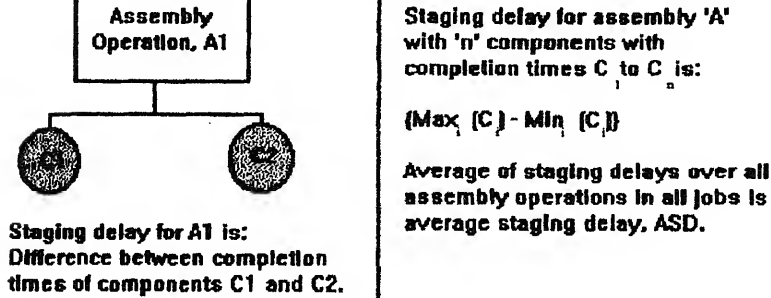
Figure 1. Staging delay in assembly jobs.

completion of other components or sub-assemblies with which it is to be assembled (figure 1).

A number of reports in the literature have attempted to devise priority dispatching rules to overcome ill effects of staging delay (Huang 1984; Adam *et al* 1987; Fry *et al* 1989). Fry *et al* (1991) also give a succinct summary of major findings in earlier works. A common theme in a majority of these works involves partial anticipation of staging delay problems ahead in some forms and consequent re-prioritising, using rather centralised strategies. In such schemes, delays in some branches of an assembly can lead to some loss of priorities for operations along other branches, in effect promoting a tendency for delays on those other branches with respect to the (original) ODDs. Such proliferation of possible delays can lead to undesirable consequences. There appears to be no dominant strategy emerging, robust with respect to depth/breadth of product structures as well as looseness or tightness of due dates etc.

Thus we find that on the one hand the MOD rule has not been studied in the context of assembly job shops, and on the other hand a variety of rules studied for assembly job shops in recent literature do not seem to have put forth a reliable approach to the problem while they attempt some anticipation of staging delays ahead in time. In our approach for controlling the phenomenon of staging delays, we employ a certain look-back strategy (rather than look-ahead), rooted in the fact that long operational tardiness of a job would normally be associated with a long waiting time at some machine for some operation on the job (possibly upstream). Before discussing the motivation and logic of our approach (in § 3), we first indicate our proposed rule form in § 2.

2. Proposed rule

Let's define a job in general as an assembly. This is represented as a directed convergent tree whose nodes represent operations and arcs indicate immediate precedence. The root of the tree is the job's final operation. An operation having more than one immediate predecessor is an assembly operation. With each immediate predecessor, say p, of an assembly operation, is associated a component of that assembly. Components can be defined as maximal linear strings of operations which are mutually exclusive and such that only the first operation in the string may be an assembly operation, e.g. in figure 3 below in § 4

the components are indicated as operation strings A, B, C and D. The notion of components is also vital to the concept of staging delay in assembly job shops, as illustrated in figure 1.

Typically, several operations on different jobs may compete for the same facility or machine at any time. An operation can qualify for competition for a machine's time only when all its predecessor operations are completed. Only when this happens, is an operation considered to be waiting for its required machine. In this context, if an operation in a job is waiting at a machine, and cannot possibly be completed within the operation's due date, then the said job is referred to as an operationally late job by the said machine.

Before defining our proposed rule, we assume that the modified operational due dates have been defined. Baker (1984) had indicated that the MOD concept, defined first for serial structures for jobs, could be usefully extended to assembly jobs. Here, we assume that job due dates have been generated based on total work content along critical path (TWKCP) as has generally been the case in literature (e.g. Adam *et al* 1987; Fry *et al* 1989 etc.). An operation's due date (ODD) is analogously generated based on the job due date and total work content along the path from the operation to job completion excluding the operation in question. Given ODDs, the MODs are defined following Baker (1984).

We propose a priority dispatching rule of the following form.

- Operationally non-late jobs to be ranked by increasing ODD (which is same as MOD here) and to form the lower priority class;

- Operationally late jobs (which are waiting since they became eligible for processing at the machine) to form the higher priority class; and jobs in this to be ranked by increasing value of the priority index:

$$(\text{Processing time})/(\text{Waiting time})^{\alpha},$$

where α is the single control parameter (positive). Mathematically, for operation i of job j we define (with reference to table A2 in the appendix for symbol definitions):

- ATi, j = arrival (eligibility) time of operation (i, j) at its required machine. Also, $\text{MOD}_{i,j}$ gets defined at any time (before starting operation i of j) at T_{now} as the maximum of $\text{ODD}_{i,j}$ and $\text{EFT}_{i,j}$ (which is equal to $T_{\text{now}} + t_{i,j}$).

We shall refer to our rule as MODWIP – modified operational due date rule incorporating waiting time induced prioritization (figure 2). Extreme values of α, viz. zero and infinity, impart SPT and FCFS priorities respectively to operationally late jobs. Incidentally,

Thus,

Priority Index $(\text{PI})_{i,j} = \{t_{i,j} / (T_{\text{now}} - AT_{i,j})^{\alpha}\}$,   if, (i, j) is operationally late

$\qquad\qquad\qquad = \text{ODD}_{i,j}$,   otherwise.

$(0 \leq \alpha \leq \infty)$

Figure 2.   Modification of MOD using waiting induced acceleration.

$= -\infty$ would correspond to LCFS priority; but we are not interested in $\alpha < 0$. Some
alysis concerning the role of α is given in § 6 later.

This type of interpolation between SPT and FCFS does not appear to have been studied
the literature, although well known rules in literature include:

) Truncation of SPT by FCFS (among jobs with waiting time exceeding a limit) which
forms a higher priority class;

) truncation of FCFS by SPT (when queue length at a machine exceeds a limit).

We sought a more elegant way of interpolating between SPT and FCFS than truncation
SPT by FCFS and vice versa. At one extreme we know that SPT minimises F_{avg} and
$_{avg}$ and at the other extreme, it is easy to see the following:

mma 1. In static single server scheduling FCFS minimises F_{max}.
his can be seen as a corollary of the result that EDD minimises L_{max}; by defining the
ɔ due dates as the job arrival times themselves.]

We have considered the FCFS rule as a limiting member of the family of the weighted-
'T rules, with priority indices $\{p/w^{\alpha}\}$ as α goes to infinity. It turns out, as we shall see in
ɔ, that this interpolation has interesting consequences; including the fact that as α goes
infinity the behaviour of the rule comes to an essential agreement with the behaviour
FCFS, in the case where waiting times are very large as compared to the processing
nes.

Comprehensive surveys of dispatching priority rules are given by Conway *et al* (1967);
ɔre (1966); Day & Hottenstein (1970); Panwalkar & Iskander (1977); Blackstone *et al*
982) and Haupt (1989). For a comprehensive account of fundamental rules and algo-
hms in sequencing and scheduling, readers may refer to Lawler *et al* (1993).

Having defined our rule form, we now turn to some aspects of its motivation.

Preliminary motivation for using waiting times in dynamic prioritization

simple motivation for our rule was as follows. Among the operationally late jobs at a
achine, the ones likely to suffer large tardiness eventually, under the MOD rule, include
ɔ rather long jobs (long operation on the machine). Such long jobs would suffer a con-
ued discrimination under SPT priority (among operationally late jobs). Our rule gives a
dressal to such long and operationally late jobs, conditional upon their waiting time, so
to curtail the long tail of the tardiness distribution. The concern for the long tail here is
ɔviously important in controlling staging delay (where any assembly operation waits for
rival of all of its components).

The use of waiting time in a dispatching priority index may be said to incorporate a
nple look-back feature, which perhaps may appear to include redundant information if
ɹe were to look at the overall control decision problem in the framework of a Markovian
ɔcision Process. Yet the fact remains that among the class of reasonably simply defined
spatching priority rules, the use of look-back information in a rule cannot be readily

effectively in the pioneering works due to Jackson and others (referred to in Conway *et al* 1967 and Kleinrock (1975)); where priority index was an affine function of waiting time. Moreover, such look-back was used precisely out of concern for "the upper tail of the waiting time distribution" (p. 180, Conway *et al* 1967). In our proposed rule form, we seek a weighted-SPT type priority index and use exponentiation of waiting time in our index, with a single control parameter. Concerning the choice of α in our rule form, we give some exact and heuristic analysis in § § 6 and 7. For the bulk of our experimentation, we used $\alpha = 0.5$; and the results of this, in comparison with several well-known priority dispatching rules are given in § 4 below.

4. Dispatching priority rules used in the experiment

While considering dispatching priority for an operation in an assembly job, if we keep in view only the (unique) directed path from the said operation to the root node of the concerned assembly tree, then conventional job shop (non-assembly) priority dispatching rules are immediately applicable (as default options).

On the other hand, an intrinsic complication in assembly job shop scheduling exists as follows. In defining dispatching priority rules, it would be desirable to know to what extent the waiting time encountered along one path to the root node of an assembly tree has correlation with the waiting time encountered along other similar paths; while any such correlation generally cannot be independent of the priority dispatching rules used at various machines. In fact, reliable anticipation of waiting times concerns transient analysis of a (dynamic priority) queueing network, which is practically out of question. Therefore,

Table 1. Characteristics of well-known priority rules used for comparison in experimentation.

Rule	Remarks
TWK	Consistent priority across the machines; lowers MFT & MTD; long jobs face continual discrimination, so long tail of tardiness distribution
EDD	Consistent priority across the machines; does not recognise operational lateness; not effective with respect to MTD, if due dates are tight
FCFS	Discourages job passing (overtaking), thereby avoiding high variability in flow time and tardiness; in a single machine static scheduling, FCFS minimises F_{max} (exactly as EDD minimises L_{max}) (a simple to derive, but not explicitly referred fact, which throws light on the behaviour of our rule, MODWIP, when α goes to infinity)
MSLK	Slack is difference between due date and earliest finish time; jobs with lower slack get higher priority; protects long jobs from excessive delays; not effective at lowering mean tardiness; has certain anti-SPT behaviour (Baker 1984)
MOD	Maximum of ODD and EFT for that operation; effective in minimising the conditional mean tardiness due to SPT effect, when jobs are operationally late, but long tail of tardiness distribution not controlled
TWK-IR	Refer figure 3. Priority is based on min TWK and ties broken using max importance ratio; tries to maintain parity in progress; given two components of an assembly with long strings of operations, progress through operations along one branch gets de-prioritised unless requisite progress occurs along the other branch (component) and vice versa; some instability in this local control of staging delay can affect job tardiness.

en defining a simple yet sound scheme for updating operational due dates, is not an easy
oblem in an assembly job shop.

In view of the two reasons cited above, it is not surprising that generally in testing
patching rules for assembly job shops, a majority of contending rules are from conven-
nal job shop literature; as rather little seems to be known about robust dispatching rules
assembly job shops.

Given below is a list of main basic factors or statistics which are very widely used as
lding blocks in dispatching rules in job shops (as borne out by well-known surveys such
Panwalkar & Iskander (1977)) and a set of rules related to each basic factor.

Work content,	SPT, LPT, MWKR, LWKR.
Due date,	EDD, ODD, MOD.
Arrival time,	FCFS, LCFS.
Operational milestones,	ODD, MOD, OSL.
Slack,	MSLK, OSL.
Parameter ratios.	CR, COVERT.

For comparing our proposed rule, we have used certain rules from this list and also some
omising rules from assembly job shop scheduling literature, e.g. Fry *et al* (1989, 1991).
e rule given by Adam *et al* (1987) was neither considered by Fry *et al* (1989), nor here,
it leads to a biased account of the unfinished work along other parallel branches in a job,
d more so when there are many levels of assembly. Rules considered here are described
able A1 of the appendix and certain relevant observations are given in table 1.

RNOP = Remaining number of operations along that branch till end-product level.

RNOP for D = 3,
RNOP for C = 2,
RNOP for B = 4,

So, RNOP_MAX = 4 (Along B)

IR=RNOP/RNOP_MAX

Currently,

IR for D = 3/4 = 0.75
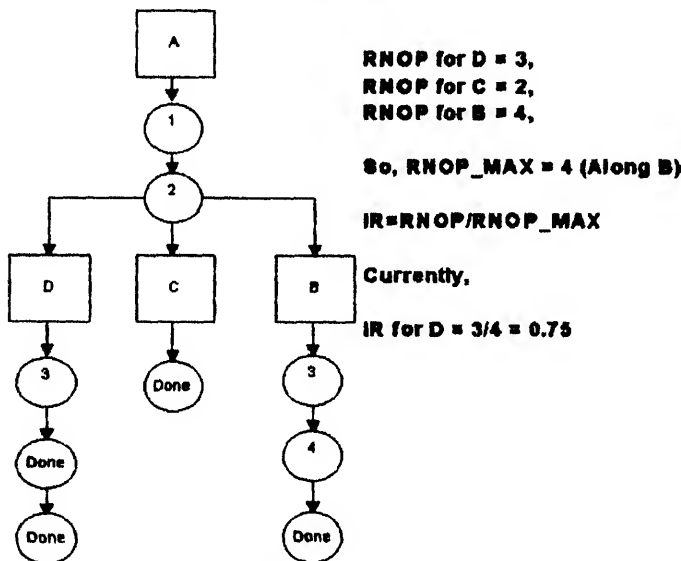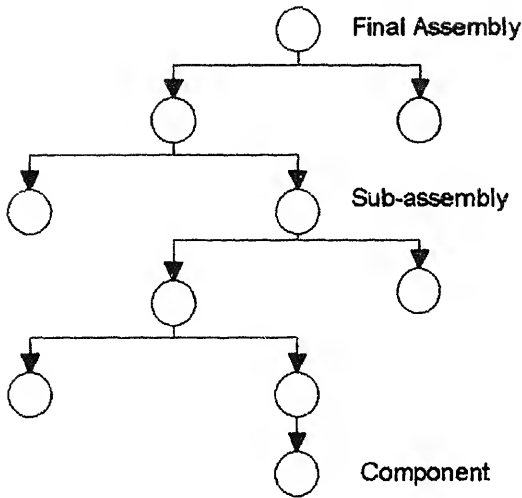
Figure 4. Tall product structure.

5. Numerical data and results

In our experimentation, we were interested, among other things, in the effect of product structure on the behaviour of different scheduling rules. Therefore, a common set of three product structures, representative of tall, mixed and flat structures, was used for all scheduling rules. These are illustrated in figures 4, 5 and 6. The other data on assembly jobs were generated probabilistically as is usually done in related literature. Table 2 gives the complete scheme for the generation of random jobs. Some elaboration follows the table.

We follow the assumption of Poisson job arrivals and exponential processing times as is the standard practice in literature (e.g. Baker & Kanet 1983, Baker 1984). Poisson job arrivals imply a memoryless arrival process so that there is no question of use of, or bias due to, information on past or future arrivals (in expectation) or in choice of scheduling decisions. Choice of exponential processing time distribution is helpful here as it exhibits
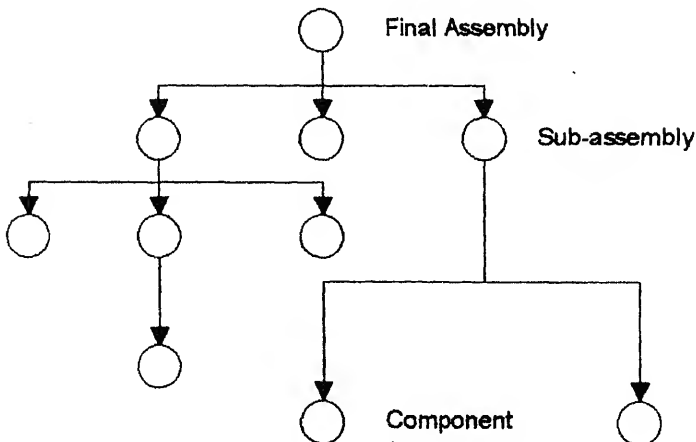
Figure 5. Mix product structure.
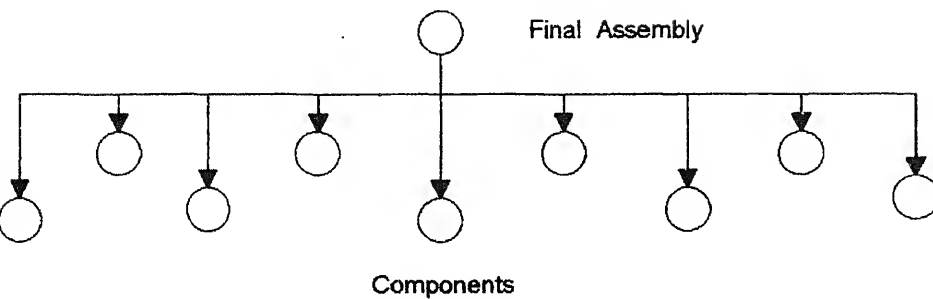
Figure 6. Flat product structure

ong tail. The relative goodness of priority dispatching rules is not likely to be affected
choice of processing time distribution as long as ample variability in processing times
l sufficient queueing exist in the system.

The values of mean processing time (5 h) and mean inter-arrival rate (12 h) were fixed
ed on an intended aggregate shop loading or utilisation around 0.6. Staging delay for
nponents would contribute significantly to components waiting in the system, in ad-
on to the waiting of components for a machine after they become scheduleable. It
s found that system utilisation around 60% still caused $O(10^2)$ components waiting
he system on the whole, and this level of congestion in the system was adequate for
nging out differences among different dispatching rules without causing extremely long
oilisation periods for the system.

Tables 3, 4 and 5 give numerical results for the fourteen dispatching rules tested in
pect of the six performance criteria. The details of the notations are given in tables A1
l A2.

le **2.** Simulation set-up.

ameter	Value	Remark
nber of machines	20	Non-identical
duct structures	3	Refer to figures 5, 6 & 7
nber of components per end product	10	Includes sub-assemblies
tribution for number of operations per nponent [range]	Uniform [1,5]	
tribution for routing of a component among chines [range]	Uniform [1,20]	
cess of job arrivals [mean inter-arrival time]	Poisson [12 h]	
tribution for proc. time for any oprn. [mean]	Exponential [5 h]	Includes set-up times
date tightness factor	6	DD = arvl + proc*6
gth of each simulation run	12000	End products
-in period length	2000	End products
rval (period) between successive instants of tem statistics collection	8 h	
ue of α	0.5	Refer to §§ 6 and 7.

Table 3. Results for tall product structure.

Rule	MFT	MTD	CMT	NTJ %	NINTN	AS
S/OPN	458.86	9.01	**72.86**	0.123	**332**	158
TWK-FIFO	467.33	19.47	**109.74**	0.177	**464**	160
TWK-IR	476.39	21.82	**115.37**	0.189	**490**	154
IR-S/OPN	463.91	16.44	**97.02**	0.169	**381**	142
IR-TWK	458.09	15.54	**99.16**	0.156	**382**	140
MS-IR	456.84	17.70	**111.08**	0.159	**491**	153
MS-TWK	478.16	24.00	**123.02**	0.195	**493**	159
TWK	419.56	6.53	**83.63**	0.078	**351**	150
EDD	463.4	8.04	**68.75**	0.117	**335**	163
FCFS	458.14	6.78	**73.15**	0.092	**326**	155
MSLK	447.61	11.53	**90.54**	0.125	**357**	159
MODWIP	**455.85**	**5.58**	**64.75**	**0.086**	307	**156**
RAN	487.11	28.62	**130.11**	0.107	**545**	171
MOD	460.03	8.33	**78.81**	0.105	**303**	159

6. Regarding choice of α

Since our dispatching priority rule form p/w^{α}, where p is processing time and w is waiting time undergone, relates to the SPT/C or weighted SPT rule (Conway *et al* 1967; Baker 1974), we first explore certain implications of this analogy.

Consider first an M/G/1 queue with static priority classes (Conway *et al* 1967; Jaiswal 1968; Klienrock 1975). For class 'k' jobs, assume there is a positive constant c_k representing a weighting factor for flow times of class k and p_k is a constant processing time for jobs in class k. It is known that in this situation with static non-preemptive priorities among classes based on

$$\text{Min}_k(p_k/c_k)$$

Table 4. Results for mixed product structure.

Rule	MFT	MTD	CMT	NTJ %	NINTN	AS
S/OPN	431.50	24.64	**77.92**	0.316	**374**	14
TWK-FIFO	428.21	36.55	**110.74**	0.330	**522**	14
TWK-IR.	445.17	42.81	**114.04**	0.375	**486**	13
IR-S/OPN	431.72	33.46	**97.33**	0.343	**389**	12
IR-TWK	439.69	33.06	**96.13**	0.344	**416**	12
MS-IR	421.59	32.59	**102.54**	0.317	**442**	13
MS-TWK	438.98	38.04	**104.02**	0.365	**430**	13
TWK	405.73	21.26	**85.92**	0.247	**371**	14
EDD	424.88	23.59	**74.75**	0.315	**335**	14
FCFS	422.34	19.04	**71.35**	0.266	**289**	13
MSLK	412.70	24.87	**89.85**	0.276	**393**	13

Table 5. Results for flat product structure.

Rule	MFT	MTD	CMT	NTJ %	NINTN	ASD
S/OPN	421.55	138.83	166.98	0.831	607	274.7
TWK-FIFO	441.42	162.42	199.70	0.813	661	267.8
TWK-IR	414.50	134.80	168.60	0.799	564	254.9
IR-S/OPN	410.01	129.88	161.23	0.805	557	249.3
IR-TWK	406.89	127.79	161.25	0.792	538	249.7
MS-IR	393.68	118.91	159.76	0.744	596	248.5
MS-TWK	429.60	148.76	181.41	0.820	619	263.0
TWK	389.17	115.04	156.77	0.733	582	245.4
EDD	382.10	106.53	140.56	0.757	502	258.1
FCFS	396.88	115.61	145.33	0.795	569	257.8
MSLK	390.94	113.66	148.96	0.763	527	242.9
MODWIP	383.83	103.75	135.25	0.767	482	252.4
RAN	495.10	198.42	255.20	0.853	673	277.3
MOD	386.90	106.90	137.57	0.777	528	249.7

and FCFS within each job class, (1) is equivalent to

$$\text{Min } E\left\{\sum_{i \in I} c_i\right\}, \tag{2}$$

where the sum extends over jobs i waiting in the system (at a random point in time). Applying Little's law we also know that (2) is equivalent to

$$\text{Min }\left\{\sum_k \lambda_k \cdot w_k \cdot c_k\right\}, \tag{3}$$

where w_k denotes mean waiting time (in queue) for jobs in class k, since the average number of jobs in the machine is invariant as there is no avoidable machine idling.

Interpreting the sum in (3) as the time rate of accumulation of system cost defined appropriately, it is seen that (3) is also equivalent to

$$\text{Min } E\{w_R \cdot c_R\}, \tag{4}$$

where, R denotes the class index of a random job coming to the system and W_R is a random sample of a waiting time within the jobs of class R.

Now, let's consider our rule form

$$\text{Min } \{p/w^\alpha\}. \tag{5}$$

Here, essentially the priority 'class' of a job changes over time while the job waits, as it depends on both p as well as w (which increases). Therefore strictly speaking it is not obvious whether, analogous to (4), we would have

$$\text{Min } E\{W_R \cdot (W_R)^\alpha\}, \tag{6}$$

if we use dispatching rule as in (5) and denote W_R a random sample of waiting times of jobs among class R, where the class R itself is drawn at random with probability

$\lambda_k/(\Sigma_j \lambda_j)$ of R being k. Writing the said random sample W_R simply as W, let us write (6) as,

$$\text{Min } E(W^{(1+\alpha)}). \tag{7}$$

Conjecture. In an M/G/1 queueing system with heavy congestion so that with overwhelming probability,

$$W \gg p, \tag{8}$$

the dynamic dispatching rule in (5) would nearly (asymptotically, in terms of ratio of objectives) achieve the objective in (7).

The reasons why we have made a statement only in the form of a conjecture are as follows.

Reason 1.   The equivalence of (5) and (7) is not expected to hold if asymptotic condition of (8) does not hold. This is explained using proposition 1 below.

PROPOSITION 1
Consider a **static** *situation where jobs* $j = 1, 2, \ldots, n$ *have already undergone waiting times* w_j *respectively. Suppose the objective is to*

$$\text{Min } \sum_j (W_j^{(1+\alpha)}), \tag{9}$$

where, W_j *denotes the total waiting time of job 'j' just before it gets served. If,* p_j *denotes processing time of job 'j', then priority rule in (5) asymptotically achieves objective in (9), if,* $W_j \gg p_j$ *for all j.*

Proof of proposition 1.   The proof follows simply from the usual pair-wise interchange perturbation applied to a sequence of jobs. e.g. job 1 would immediately precede job 2 in an optimal permutation only if:

$$p_2 \cdot \text{d}/\text{d}x\{(W_1 + x)^{(1+\alpha)}\}_{\text{at } x=0} \geq p_1 \cdot \text{d}/\text{d}x\{(W_2 + x)^{(1+\alpha)}\}_{\text{at } x=0}, \tag{10}$$

$$\text{i.e. } p_2 \cdot W_1^{\alpha} \geq p_1 \cdot W_2^{\alpha},$$

$$\text{i.e. } p_1/W_1^{\alpha} \leq p_2/W_2^{\alpha}, \tag{11}$$

which yields the proof.

Note that in the absence of the asymptotic condition , essentially, (8) or (10) would not hold and nor would (11). That is why the objective in (7) cannot be expected to strictly follow from the rule form of (5).

Reason 2.   The linking behaviour of total differences in value of potential function for Markov chains, with costs in more than one dimension and with non-homogeneous in-

renewal process, the limiting behaviour of local differences in the renewal function defined on the linear state-space). In fact there exist well-known open conjectures concerning asymptotic behaviour in queueing systems (e.g. the conjecture of Jackson in Conway *et al* 1967, p. 181) for which the standard limit theorems normally invoked in the context of Markov chains or Markov renewal processes do not seem to suffice. In fact the only known (to us) proof of optimality of the SPT/C rule in a truly dynamic scheme of priorities (not among classes of jobs but among individual jobs present in any given system state) is due to Kakalik & Little (1969) which exploits in great depth the duality between LP and MDP, and no comparable studies for more general objective function seem to be known.

It may be mentioned here that as α goes to ∞, what proposition 1 tells us, in the case where $p \ll W$, stands in essential agreement with the fact seen earlier that FCFS minimises $F_{\max}$. This is because in that case minimising $W_{\max}$ is practically as good as minimising $F_{\max}$ from the user's point of view.

Note. In the study of dynamic dispatching rules and their optimisation at least in an approximate sense, the use of Little's law is found to shed some light on the nature of optimisation as seen from a queueing control theoretic point of view. This has also been seen, in the context of another type of problem, by Awate & Neeraj Kumar (1992).

7. Empirical guideline regarding choice of α

During early stages of our experimentation, and before our analysis given in § 6, we had considered α as powers of 2 [0.5, 1, 2, 4, 8, 16]. We found values of α greater than one to be not promising enough to warrant further consideration. Here we suggest a heuristic explanation for this.

From a practical point of view, we seek reasonably low mean flow time, mean tardiness, conditional mean tardiness and a small value for (say) 99 percentile point of tardiness distribution (to indicate avoidance of a long tail). Also, rather than seeking to control the 99 percentile point, it could suffice to 'keep in control' the mean and variability of tardiness. Although variance of tardiness is not directly captured by any (non-central) moment of tardiness distribution, keeping in control the first two (first and second) moments of tardiness would be a reasonable objective. Now applying the heuristic arguments concerning our conjectured relation between rule in (5) and effect in (7), we are led to the choices of $(1 + \alpha)$ as 1 and 2; i.e. $\alpha = 0$ and $\alpha = 1$.

In light of the facts and analysis concerning our objectives, we are led to a guideline for range of α between 0 and 1.

Table 6. Results for tall product structure – Variation in α.

α	MFT	MTD	CMT	NTJ %	NINTN	ASD
0.00	468.46	10.29	76.89	0.133	390	166.8
0.50	455.85	5.58	64.75	0.086	307	156.8
1.00	468.20	10.06	81.68	0.123	394	164.9

Table 7.   Results for mixed product structure – Variation in α.

α	MFT	MTD	CMT	NTJ %	NINTN	ASD
0.00	413.13	19.84	75.56	0.248	291	142.1
0.50	415.82	15.80	65.17	0.242	274	139.1
1.00	424.11	21.66	71.56	0.268	291	139.0

Our current experimentation tests values of α as 0, 0.5 and 1. Among these, $\alpha = 0.5$ performed better. We did not seek a fine tuning of value of α because we expect, based on the analysis above and numerical results, that the value of α recommended should be a simple or robust one. As noted by Baker (1974), the SPT effect itself often has unexpected side benefits, such as considerably reducing the mean of lateness, so that the concern for $\alpha = 0$ case remains strong. On the other hand choosing $\alpha = 1$ corresponds to a quadratic cost of waiting in (7). Thus, $\alpha = 0.5$ seems a simple and robust compromise. Results for different values of α are given in tables 6, 7 and 8.

8.   Conclusions

Use of job waiting times at a machine in the dispatching priority index for operationally late jobs has been found to provide improvement in terms of tardiness measures: mean tardiness, conditional mean tardiness and 99 percentile point [1] of tardiness distribution. In case of flat product structure, we got more significant improvement in 99 percentile point of tardiness distribution; here the fraction of jobs tardy was high (more than 75%)*. Here, jobs in the upper tail of tardiness distribution were in relatively large numbers, and could be expected to be among operationally late jobs and consequently influenced by our rule's discrimination based on waiting times. On the other hand, in case of tall product structures, the fraction of tardy jobs was small (of the order of 10%). Correspondingly here, our rule, based on discrimination among operationally late jobs, got invoked less often. The improvement we got was relatively more significant in mean tardiness indicating that some operationally late jobs, probably upstream, were accelerated by our rule and thereby prevented from being tardy at the end. In case of mixed product structure, we got significant improvement in mean tardiness, conditional mean tardiness as well as 99 percentile point of tardiness distribution.

Table 8.   Results for flat product structure – Variation in α.

α	MFT	MTD	CMT	NTJ %	NINTN	ASD
0.00	390.73	112.67	146.10	0.799	560	256.3
0.50	383.83	103.75	135.25	0.767	482	252.4
1.00	392.68	109.34	141.50	0.785	544	255.1

---

[1] For the allowance setting mechanism, namely, six times the processing time does not provide for extra allowance in case of flat structure which is exposed to higher staging delay; which leads to higher fraction of tardy jobs as

As we had indicated earlier in the paper and in conformity with some of the earlier works cited from the literature, the SPT effect seems to play a crucial role in tall product structure and indeed we found the MOD rule to be a very strong competitor, especially in tall product structure. Our current results indicate the MOD rule giving a slightly better value for the 99 percentile point of the tardiness distribution (but not for MTD and CMT). In our further experimentation, involving variation of more number of factors, we intend to take a closer look at the cited phenomenon in tall product structures.

Regarding choice of parameter 'α' in the experimentation of waiting time in our priority index, our results indicate that $\alpha = 0.5$ is better than $\alpha = 0$ (equivalent to SPT among operationally late jobs) and $\alpha = 1$ (equivalent to FCFS among operationally late jobs). Based on certain analytical relations between queueing control problems and priority dispatching problems, we have indicated earlier the desirability of α lying between 0 and 1; reflecting concern for the mean and the second moment of tardiness. Based on our current findings, $\alpha = 0.5$ is consistently better at tardiness measures as seen in tables 6, 7 and 8. Further study of the role of value of α will be part of future research.

Our new proposed rule, developed in the context of assembly job shop scheduling, gives an effective and efficient extension of the well-known MOD rule defined in the literature for job shop scheduling with due dates. On the other hand, we have derived a practically very interesting interpolation between SPT and FCFS dispatching rules and delineated its relation to a rich analogy with a queueing control problem. This also seems to warrant further study from the point of view of job shop scheduling in general.

Our findings are also essentially in agreement with some previous ones in literature that the most complicated rules in job shop scheduling are hardly the most effective ones and simple-looking rules based on sound analysis do have scope for furthering improvements.

Appendix A.

Table A1. Rules used in simulation.

Name	Explanation		Formula	
	Primary rule	Tie-breaker	Primary rule	Tie-breaker
S/OPN	Slack per operation	—	$\text{Min}_i[(d_i \cdot \text{EFT}_i)/\text{RNOP}_i]$	—
TWK-FIFO	Total work content	First in, first out	$\text{Max}[\Sigma\, t_i] : i = \text{current}$ oprn* to end oprn	$\text{Min}_i[r_i]$
TWK-IR	Total work content	Importance ratio	$\text{Max}[\Sigma\, t_i] : i = \text{current}$ oprn to end oprn	Refer to figure 3
IR-S/OPN	Importance ratio	Slack per operation	Refer to figure 3	$\text{Min}_i[(d_i \cdot c_i)/\text{RNOP}_i]$
IR-TWK	Importance ratio	Total work content	Refer to figure 3	$\text{Max}[\Sigma\, t_i] : i = \text{current}$ oprn to root oprn
MS-IR	Modified slack	Total work content	Max. IR, if SL < 0 Min. SL, o/w	$\text{Max}[\Sigma\, t_i] : i = \text{current}$ oprn to root oprn
MS-TWK	Modified slack	Importance ratio	Min. TWK, if SL > 0 Min. SL, o/w	Refer to figure 3
TWK	Total work content	–	$\text{Min}_i(t_i)$	–
EDD	Earliest due date	–	$\text{Min}_i(d_i)$	–
FCFS	First come first served	–	$\text{Min}_i(r_i)$	–
MSLK	Min. slack	–	$\text{Min}_i(d_i \cdot \text{EFT}_i)$	–
MODWIP	Refer to § 2	–	Refer to figure 2	–
RAN	Random	–	$U(0,100)$	–
MOD	Modified operational due date	–	$\text{Max}[\text{ODD}, (T_{\text{now}} + t_{i,j})]$	–

* Oprn – Operation.

Table A2. Symbols and explanations.

Symbol	Explanation	Formula
i	Operation	—
j	Job	—
k	Flow allowance factor	6
n	Number of jobs (end products)	—
α	Control parameter for MODWIP	0.5
t_j	Total processing time of job j	$\Sigma\, t_{i,j}$ over all i for each j
d_j	Due date of job j	$r_j + k^* t_j$
r_j	Arrival time of job j	—
EFT_j	Earliest finish time of job j	—
c_j	Completion time of job j	—
$t_{i,j}$	Processing time of oprn i of job j	—
T_{now}	Current time	—
RNOP_j	Remaining number of oprns of job j	—

Table A2. (Continued)

Symbol	Explanation	Formula
$ODD_{i,j}$	Operational due date of oprn i of job j	$ODD_{i-1,j} + [t_{i,j}/t_j]^*(d_j - r_j)$ where, $ODD_{0,j} = r_j$
$AT_{i,j}$	Arrival time of oprn i of job j	–
$SL_{i,j}$	Modified slack for oprn i of job j	$d_j - T_{now} - \Sigma\, t_{i,j}$ (over operations following i, j)
MFT	Mean flow time	$\{\Sigma\,(c_j - r_j)\}/n$
T_j	Tardiness of job j	max (lateness, 0) for job j
MTD	Mean tardiness	$\{\Sigma\,(T_j)\}/n$
CMT	Conditional mean tardiness	$\{\Sigma\,(T_j)\}/n_t$
NTJ	Percentage of number of tardy jobs	n_t/n [n_t: all jobs such that $c_j > d_j$]
NINTN	99% point of tardiness distribution	Refer to figure A1
ASD	Average staging delay	Refer to figure 1
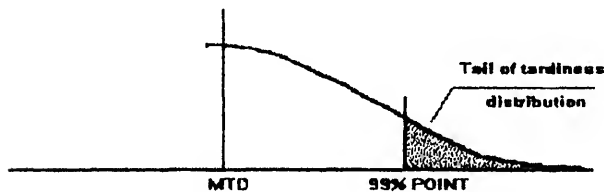
Figure A1. 99% Point of tardiness distribution statistics.

References

Adam N R, Bertrand J W, Surkis J 1987 Heuristics for assembly job shop scheduling. *IIE Trans.* 19: 317–328

Awate P G, Neeraj Kumar 1992 Development of adaptive control strategies for dispatching rules in dedicated FMS with random arrivals. Paper presented at ORSI Annual Convention at the Indian Inst. Manage., Ahmedabad

Baker K R 1974 *Introduction to sequencing and scheduling* (New York: John Wiley & Sons)

Baker K R 1984 Sequencing rules and due date assignments in a job shop. *Manage. Sci.* 30: 1093–1104

Baker K R, Kanet J J 1983 Job shop scheduling with modified due date. *J. Oper. Manage.* 4: 11–22

Blackstone J H, Phillips D T, Hogg G L 1982 A state-of-the-art survey of dispatching rules for manufacturing job shop operations. *Int. J. Prod. Res.* 20: 27–45

Conway R W, Maxwell W L, Miller L W 1967 *Theory of scheduling* (Reading, MA: Addison-Wesley)

Day J E, Hottenstein M P 1970 Review of sequencing research. *Nav. Res. Logist. Q.* 17: 11–39

Fry T D, Philipoom P R, Russell R S 1983 Due date assignment in a multi-stage job shop. *IIE Trans.* 21: 153–161

Fry T D, Philipoom P R, Russell R S 1989 Dispatching rules in multi-stage assembly shops. *Int. J. Prod. Res.* 27: 671–686

Fry T D, Philipoom P R, Russell R S 1991 A preliminary investigation of multi-attribute based sequencing rules for assembly shops. *Int. J. Prod. Res.* 29: 739–753

Gere W S 1966 Heuristics in job shop scheduling. *Manage. Sci.* 13: 167–190

Haupt R 1989 A survey of priority rule-based scheduling. *Oper. Res. Spektrum* 11: 3–16

Huang P Y 1984 A comparative study of dispatching rules in a hybrid assembly/job shop. *Int. J. Prod. Res.* 22: 375–387

Jaiswal N K 1968 *Priority queues* (New York: Academic Press)

Kakalik J S, Little J D C 1969 Dynamic operating policies for a single server facility. Technical Report No. 47, Operations Research Centre, Massachusetts Institute of Technology, Cambridge, MA

Kleinrock L 1975 *Queuing systems. Vol II – Computer applications* (New York: John Wiley & Sons)

Lawler E L, Lenstra J K, Rinnooy Kan A H G, Schmoys D B 1993 Sequencing and scheduling: Algorithms and complexity. In *Logistics of production and inventory: Handbooks in operations research and management science* (eds) S C Graves *et al* (Amsterdam: North-Holland) vol. 4, ch. 9

Panwalkar S S, Iskander W 1977 A survey of scheduling rules. *Oper. Res.* 25: 45–61

Russell R S, Taylor B W 1985 An evaluation of sequencing rules for an assembly shop. *Decision Sci.* 16: 196–212

Wein L M, Chevalier P B 1992 A broader view of the job-shop scheduling problem. *Manage. Sci.* 38: 1018–1033

Modelling and simulation of Just-In-Time flexible systems

K RAVI RAJU[1], K RAMA BHUPAL REDDY[1] and
O V KRISHNAIAH CHETTY[2]

[1]Department of Mechanical Engineering, Indian Institute of Technology,
Hauz Khas, New Delhi 110 016, India
[2]Department of Mechanical Engineering, Indian Institute of Technology,
Madras 600 036, India
e-mail: [raviraju,krbreddy]@mech.iitd.ernet.in; ovk@acer.iitm.ernet.in

Abstract. The increasing worldwide competition requires economical manufacture, high quality and short delivery time. The Just-In-Time (JIT) philosophy of manufacturing is increasingly being considered by manufacturing organizations, as a response to the increased pressure to supply high quality products with short delivery times and at low cost. A very simple shop floor control system that was developed by Toyota in the 1970s specifically for their Just-In-Time assembly plant, has received considerable attention in the Western World, and is known as the 'kanban' system (literally translated as 'card' system). Japan's success has prompted many scholars and practitioners, to turn their attention to Japanese management practices. This article is to address the modelling, simulation and implementation issues of Just-In-Time in flexible manufacturing environments. Priority nets are used for modelling and analysis of the kanban system. A large number of simulation runs are conducted/presented to probe the behaviour of the system, with respect to different parameter changes.

Keywords. Just-In-Time; flexible manufacturing systems; kanban.

Introduction

Just-In-Time (JIT) is considered to be one of the major factors contributing to the success of Japanese industries (Sugimori *et al* 1977; Monden 1981, 1983; Schonberger 1982). The JIT approach to production was originated by Toyota in 1970s in their car assembly plants in Japan, and has found most success in repetitive manufacturing environments. The JIT philosophy has attracted increased attention from researchers who have tried to explain JIT implementation successes and failures outside Japan (Narendar *et al* 1995). The objectives of the JIT system are to maximize the velocity of material flow through the plant, by arranging materials to arrive at each stage of manufacture just in time, to be processed and move on to the next stage. JIT manufacture is a well-proven manufacturing management philosophy which aims at eliminating the Work-In-Process (WIP) and optimizes material

101

delivery timings. Traditionally it is applied in the mass manufacturing environment. In recent years, it also being implemented in flexible manufacturing environment. A survey by Kim & Schniederjans (1990) which includes 450 US manufacturing firms emphasizes that a computer-integrated JIT system is more productive, cost efficient and highly effective in producing a quality product than a JIT or CIM system alone. To date, there is a lot of reported research in the areas of JIT and flexible manufacturing individually. However, there is no detailed investigation of implementing JIT in flexible manufacturing environments. Hence, to realize the full benefits of JIT in flexible manufacturing, many implementation issues have to be investigated.

Though, JIT offers many advantages such as WIP reduction, increased flexibility and quality, reduced space requirements etc., its implementation is not always a success. Indeed, on the one hand it necessitates a specific consumption of finished products that is regular enough, while on the other, it necessitates major changes in manufacturing process such as reducing setup times, increasing reliability, improving product quality and reducing customer response time (Schonberger 1982; White 1993). Thus these problems justify the interest in methods to evaluate the performance of these systems.

Kanban system is the backbone of JIT philosophy. The philosophy behind the kanban approach is, the work should never be pushed on to the next work centre until that work centre is ready for it. This utilizes cards to authorize production and move material between work centres. In the literature, kanban systems have been studied using a variety of methods. Some researchers have analyzed JIT systems using mathematical programming (Bitran & Chang 1987; Mitra & Mitrani 1988; Wang & Wang 1991). Simulation is frequently used to evaluate JIT systems (Huang *et al* 1983; Schroer *et al* 1985; Lulu & Black 1987). System dynamics models are also used to study the kanban systems (Gupta & Gupta 1989). Mascolo *et al* (1991) have demonstrated the suitability of Petri Nets for unified modelling of kanban systems.

The objective of this paper is to establish priority nets (PRNs) as powerful and flexible tools for modelling and analysis of JIT manufacturing systems. It also details the differences between the implementation issues of conventional manufacturing systems (CMSs) and FMSs, and addresses some of the issues with petri net-based simulation.

2. JIT in flexible manufacturing environment

The success of JIT mainly depends on the nature of manufacturing environment. It is found to be more suitable for a repetitive manufacturing environment, where products are regularly ordered. The suitability of JIT for manufacturing environment is summarized in figure 1 (Voss & Harrison 1987). Systems with mid volume and mid variety (flexible manufacturing environment) almost certainly have part or all of their manufacturing suitable for JIT, those at the top left and bottom right are suitable for selected applications.

The second criterion for determining the suitability of JIT is based on the complexity of product structure and process routing, figure 2 (Voss & Harrison 1987). The more complex

VARIETY
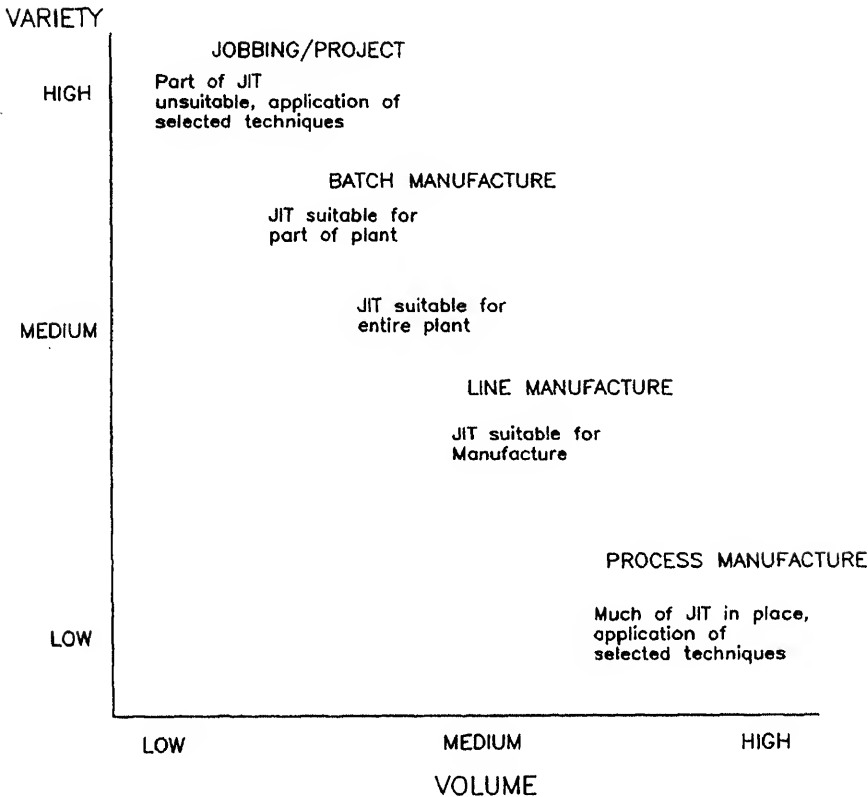
Figure 1. JIT and the manufacturing environment (Voss & Harrison 1987).

1 *JIT in FMSs*

T systems require production processes which can economically produce small runs of oducts. This flexibility is generally required to support the JIT concept. Economical nall production runs are achieved by reducing the equipment setup time which is the xed cost component associated with each production run. With the advent of FMS this quirement has become a reality. FMSs have the ability to produce a wide range of parts uickly in small lots, and thus fit well into the JIT approach. In addition, it can be stated that ithout some form of FMS or flexible manufacturing facility, it is difficult to implement T effectively within a batch manufacturing environment.

2 *JIT in flexible assembly systems*

exibility in assembly is a fast-growing prerequisite imposed by present day industry, ie to the ever-increasing variety of models, and to quickly respond to market changes. ie advent of programmable industrial robots spurred the development of FASs. JIT is iccessfully implemented in many FASs. For example, an FAS is installed at International omputers Ltd. (ICL, England) integrating assembly, testing and philosophy. This system

MRP : Manufacturing Resource Planning
PERT : Programme Evaluation and Review Technique
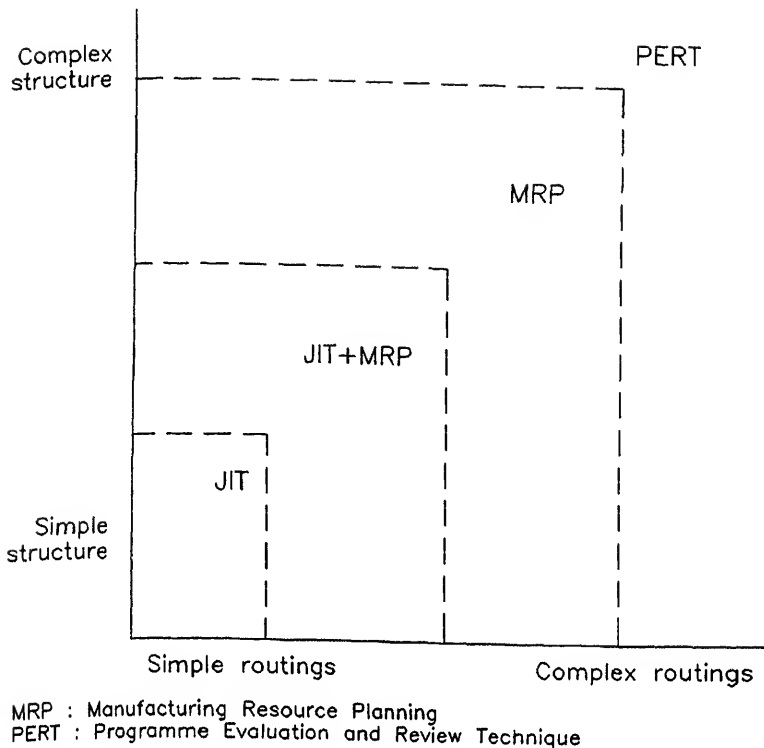
Figure 2. Suitability of JIT for controlling production (Voss & Harrison 1987).

help of central software and coded pallets (Talavase & Hannam 1988). Pye Telecommunications Ltd., successfully implemented FAS with JIT, with a 96% reduction in lead time and WIP (Wilson 1986).

2.3 *JIT in job shop environment*

JIT approach is traditionally thought to be not suitable for small industries operating as job shops. However, in a publication (Gravel & Price 1988), it is reported that Kanban method can be successfully adapted to the job shop environment. This is tested in a firm, manufacturing high quality outdoor clothing and the results of the pilot run are found to be encouraging. However, the authors felt that some of the modelling and practical issues need to be resolved before successful implementation.

3. JIT implementation issues

The implementation issues of JIT in FMSs are different from those in CMSs, since (i) the configurations, (ii) the main criteria in each of the cases are different. In the environment of FMS/FAS, the multi-functional workers are replaced by robots, conventional machines are replaced by CNC machining centres and moving carts by AGVs. In CMSs, the main criteria is to maintain the lowest possible inventories even though the utilization of the some of the machines is reduced. But in FMSs, the utilization of machining centres must

high to justify this investment. Therefore, for implementing JIT in FMSs a balance has to achieved between inventory levels and utilization levels. Also the degree of automation s to be kept low to reduce the production cost. Masuyama (1986), has discussed some the implementation issues in FMS environment.

The prominent issues for the successful implementation of JIT in either CMSs or FMSs : (1) reduction in setup times, (2) reduction in conveyance times between machines, d (3) number of kanbans and kanban sizes. However, there are additional issues to be dressed in the context of FMSs due to the fact that (a) multiple-part types are processed nultaneously, and (b) the systems have to respond to the changes in part types and lumes. The additional issues are:

kanban related issues in multi product environment,

part routings and layout, and

AGV related issues.

Reduction of setup times

onden (1981, 1983) separates setup times in CMSs into (i) internal (actions that require e machine to be stopped) and (ii) external (actions that can take place when the machine running) setup times. Efforts are then made to convert maximum possible amount of ernal setup time to external in order to reduce the overall setup time. Internal setup time FMSs is almost negligible due to flexibility of CNC machining centres with automated ol-handling facilities.
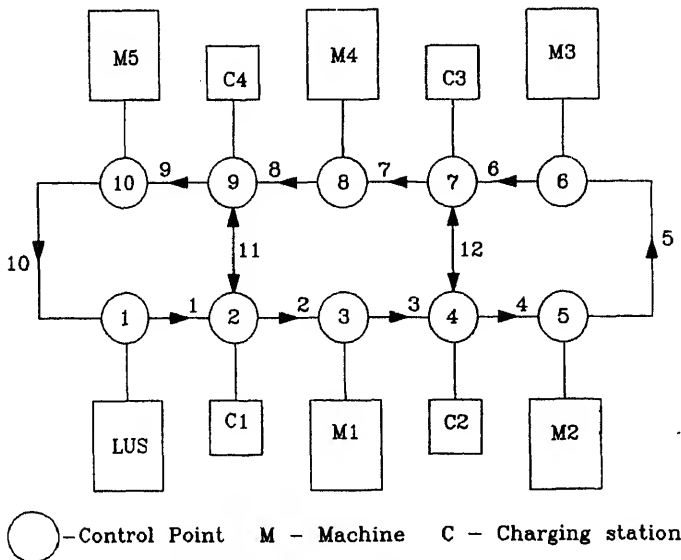
Figure 3. Layout of the system.

Table 1. Part routings and processing times.

Part type	Routing	Processing times
PR1	LUS-M1-M2/M3-M4-LUS	30-150-240/290-230-30
PR2	LUS-M4/M5-M1-M2/M3/LUS	30-205/250-170-205/290-30
PR3	LUS-M2/M3-M5-M1/M4-LUS	30-220/240-275-290/240-30

3.2 *Reduction of conveyance times*

Conveyance times in CMSs are reduced by belt conveyors, chutes or fork lifts. As these types are limited in flexibility for transferring parts/materials, they cannot be used in the environment of FMSs as asynchronous flow of material is mandatory. Owing to this AGVs are generally preferred in FMSs.

3.3 *Reduction of number of kanbans and lot sizes*

The inventory in a system is tightly controlled with the help of kanbans. A kanban system lowers WIP levels by decreasing either the number of kanbans or the kanban size (lot size). Reducing production lead time is very crucial for the operation of JIT systems with varying demand, without high levels of either safety stocks or stock-outs. Recent literature indicates that there is a significant relation between batch size and production lead time. Hence, determination of the optimum number of kanbans and lot sizes is very important in both CMSs and FMSs.

3.4 *Kanban-related issues in FMS environment*

The following issues are important in the context of FMSs.

3.4a *Pallet as a container:* In CMSs, special containers are used for each part type. In FMSs, parts are generally loaded onto pallets and routed through the machines. Pallets form the interface between the workpiece and the machines. Thus the pallets act as containers in this case. These pallets are generally coded and contain necessary information.

FMSs are capable of processing families composed of a large number of part types. Two different kinds of families (closed and open family) can be taken into account. In the first case, the mix of the part types is predetermined and the system is setup for long time production of some part types. Under these conditions, dedicated pallets with special fixtures are used for speeding up loading and unloading operations. When open families are considered, the system must quickly be able to fit every requirement. In this case adaptable pallets equipped with all-purpose clamping devices are more suitable. These pallets can

Table 2. AGV travel times.

Segment number	1 2 3 4 5 6 7 8 9 10 11 12
Travel time (s)	10 10 10 10 20 10 10 10 10 20 15 15
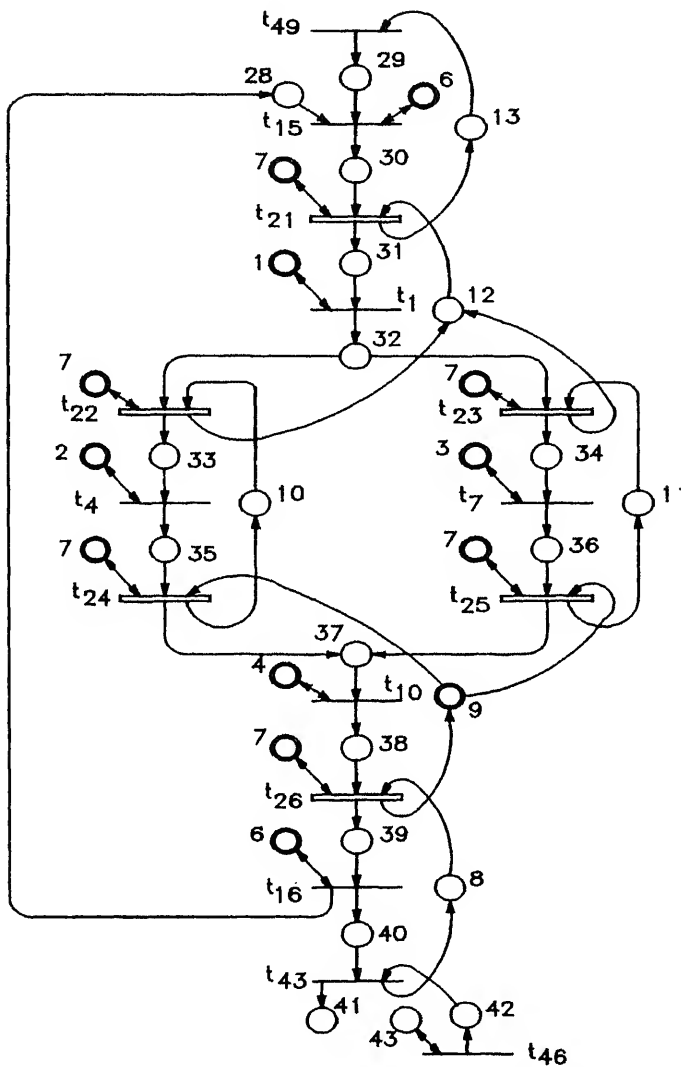
Figure 4. Priority net model for part type 1.

ceive more than one component. In one publication (Gentili 1991), it is reported that ~~MS equipped with multi-piece adaptable pallets is able to work satisfactorily in JIT~~ ~~vironment.~~

~~4b~~ *Number of kanbans:* If there are more part types in the system at a time, the ~~tal~~ inventory in the system will be very high. This is because the inventories have to ~~maintained~~ at all stages for all part types for implementing the pull mechanism. This ~~oblem~~ can be overcome by freezing schedules for a small period of time (a week or ~~days),~~ which helps in keeping the variables low, and also continuing the same part-mix ~~roughout that period.~~

Figure 5. Priority net model for part type 2.

Part type changes – Whenever a new part type enters the system, its routing and the number of kanbans have to be carefully determined. Then the 'start up' for the part type (i.e. generation of kanbans on the computers, passing the kanban information to the respective machining centres etc.) has to be planned. Similarly, whenever a part type goes out of the system, its 'cut off' has to be planned to gradually remove the inventories (kanbans) associated with the part type.

Volume changes – Even though the part mix is stable over a period of time, individual

Figure 6. Priority net model for part type 3.

3.5 *Part routings and layout*

FMSs offer flexible routing to the parts. However, simpler routings are advantageous in implementation of JIT. Adoption of Virtual Cell Manufacturing (VCM) principles and layouts is helpful for implementing JIT in flexible manufacturing environments. VCM is a new concept in small batch manufacturing (Drolet *et al* 1990). A VCM system does not promote opportunistic re-routing of each part after each operation like FMS. Instead, the route of a job (or a set of similar parts) is defined at a virtual cell. Whenever a new part type enters the system a new virtual cell is created. Drolet *et al* (1991), demonstrated that 'the VCM layout concept will function in harmony with JIT'.

Table 3. Interpretation of places.

Place	Interpretation	Place	Interpretation
1	M1 free	54	PR2 in M3's input buffer
2	M2 free	55	PR2 in M2's output buffer
3	M3 free	56	PR2 in M3's output buffer
4	M4 free	57	PR2 at LUS after processing
5	M5 free	58	Finished part of PR2 available at LUS
6	LUS free	59	Counter for PR2
7	AGV available	60	Demand for PR2
8–13	C-kanban post PR1 in kanban post	61	Place for modelling different
14–20	C-kanban post for PR2 in kanban post		demand patterns of PR2
21–27	C-kanban post for PR3 in kanban post	62	Fixtures for PR3 available
28	Fixtures for PR1 available	63	Raw material for PR3 available at LUS
29	Raw material for PR1 available	64	Loaded pallet for PR3 available at LUS
30	Loaded pallet for PR1 available at LUS	65	PR3 in M2's input buffer
31	PR1 in M1's input buffer	66	PR3 in M3's input buffer
32	PR1 in M1's output buffer	67	PR3 in M2's output buffer
33	PR1 in M2's input buffer	68	PR3 in M3's output buffer
34	PR1 in M3's input buffer	69	PR3 in M5's input buffer
35	PR1 in M2's output buffer	70	PR3 in M5's output buffer
36	PR1 in M3's output buffer	71	PR3 in M1's input buffer
37	PR1 in M4's input buffer	72	PR3 in M4's input buffer
38	PR1 in M4's output buffer	73	PR3 in M1's output buffer
39	PR1 at LUS after processing	74	PR3 in M4's output buffer
40	Finished part of PR1 available at LUS	75	PR3 at LUS after processing
41	Counter for PR1	76	Finished part of PR3 available at LUS
42	Demand for PR1	77	Counter for PR3
51	PR2 in M1's input buffer	78	Demand for PR3
52	PR2 in M1's output buffer	79	Place for modeling different
53	PR2 in M2's input buffer		demand patterns of PR3

3.6 *AGV related issues*

These are mainly, (i) the determination of number of AGVs, and (ii) dispatching. Some of the traditional dispatching rules used in push systems may not be suitable in JIT environments. In the present work, some of the implementation issues such as number of kanbans, kanban sizes and vehicle dispatching are addressed with the help of Priority nets.

4. Modelling and simulation of FMS

Modelling and analysis of FMSs implementing a pull mechanism are highly complex due to the multiple part types and AGV-based transportation environment. Priority nets can be effectively used to address these issues. To demonstrate the efficacy of PRNs for modelling pull mechanism in FMSs, a system with the following features is considered.

• The layout of the system is shown in figure 3. The system consists of five machining centres and a load–unload station connected by AGV network. It caters to a variety of part types. Three part types are considered for the present study. The machining sequences and processing times are indicated in table 1.

Table 4. Interpretation of transitions.

Transition	Interpretation	Transition	Interpretation
1	M1 processing PR1	25	Transportation of PR1, M3 to M4
2	M1 processing PR2	26	Transportation of PR1, M4 to LUS
3	M1 processing PR3	27	Transportation of PR2, LUS to M4
4	M2 processing PR1	28	Transportation of PR2, LUS to M5
5	M2 processing PR2	29	Transportation of PR2, M4 to M1
6	M2 processing PR3	30	Transportation of PR2, M5 to M1
7	M3 processing PR1	31	Transportation of PR2, M1 to M2
8	M3 processing PR2	32	Transportation of PR2, M1 to M3
9	M3 processing PR3	33	Transportation of PR2, M2 to LUS
10	M4 processing PR1	34	Transportation of PR2, M3 to LUS
11	M4 processing PR2	35	Transportation of PR3, LUS to M2
12	M4 processing PR3	36	Transportation of PR3, LUS to M3
13	M5 processing PR2	37	Transportation of PR3, M3 to M5
14	M5 processing PR3	38	Transportation of PR3, M3 to M5
15	Loading PR1 on pallet	39	Transportation of PR3, M5 to M1
16	Unloading PR1 from the pallet	40	Transportation of PR3, M5 to M4
17	Loading PR2 on pallet	41	Transportation of PR3, M1 to LUS
18	Unloading PR2 from the pallet	42	Transportation of PR3, M4 to LUS
19	Loading PR3 on pallet	43	Consumption of finished part of PR1
20	Unloading PR3 from the pallet	44	Consumption of finished part of PR2
21	Transportation of PR1, LUS to M1	45	Consumption of finished part of PR3
22	Transportation of PR1, M1 to M2	46	Arrival of demand for PR1
23	Transportation of PR1, M1 to M3	47	Arrival of demand for PR2
24	Transportation of PR1, M2 to M4	48	Arrival of demand for PR3

- The AGV track layout consists of both uni-directional and bi-directional segments. Control points are provided on the layout to avoid vehicle collisions. Bi-directional AGVs are used in the system. AGVs, when unassigned, wait in the charging stations provided on the track.

- The system uses only conveyance kanbans (C-kanbans) for implementing pull mechanism.

The following assumptions are made:

(1) An AGV takes 30 seconds for loading/unloading and 5 seconds to cross a control point. The time required to cover different segments is given in table 2.

(2) Raw materials are available in perennial supply.

4.1 Modelling

A hierarchical modelling methodology is adopted. At the higher level the pull mechanism is modelled. This is referred to as system net. At the lower level the AGVS is modelled, which is referred to as logistic net. For the system net a part variety based approach is adopted. The PRN model of pull mechanism of each part type is made separately. These PRNM are then linked by merging the common D-places to yield the system net. The PRNMs of each part type is shown in figures 4 to 6. The interpretations of places and transitions are given in tables 3 and 4 respectively. The logistic net of the system is shown in figure 7, details are given in Ravi Raju & Chetty (1993).
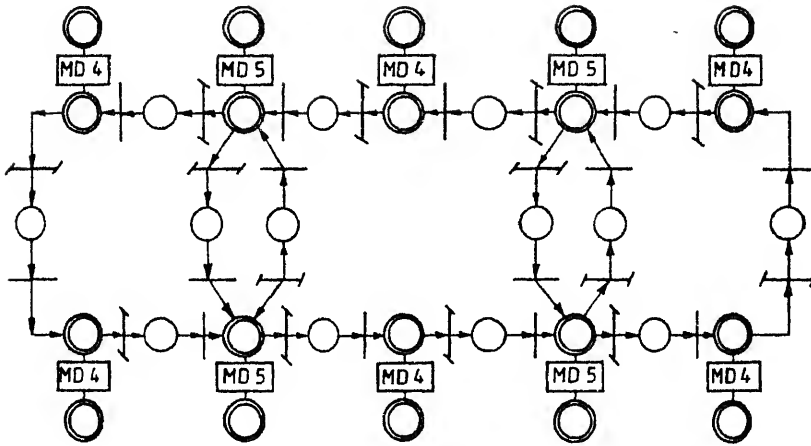
Figure 7. Logistic net of the system.

4.2 *Simulation*

The system is simulated by the concurrent execution of the system net and logistic net. Control strategies are implemented with the help of D-places. D-places associated with machines resolve conflicts among transitions representing processing activities. D-places associated with AGVs resolve conflicts among transitions representing transportation activities.

AGV movements are simulated by the logistic net, figure 7. Whenever an unload transition in the logistic net fires, it changes the attribute AGV engaged to zero, and thus terminates the firing of a transition representing a transportation activity in the system net. An Ada-based software package JIT-SIM is developed for the simulation of JIT systems. The basic inputs to the package include system net details and AGVS details. The interactive input to the package includes a number of AGVs and their locations, number of kanbans at each stage, kanban sizes, variability in the processing times, job selection rules, machine selection rules, and simulation time. In addition to the usual information, the JIT-SIM gives the important performance measures relevant to the JIT environment, such as percentage of back-ordered demand, mean tardiness, average in-process inventory, cumulative machine idle time, and overtime.

Table 5.   Inter-arrival times.

Part type	Mean arrival time	Inter-arrival times (min)
PR1	10	0 22 18 1 4 22 2 0 17 4 8 8 18 2 9 7 1 0 24 1 13 3 2 5 6 16 24 14 5 12 0 17 5 24 1 5 18 0 0 16 2 10 18 27 0 3 9 5 16 0 12 4 5 9
PR2	12	9 2 11 0 9 3 9 15 11 28 12 16 3 1 18 10 6 2 27 9 21 2 3 25 4 0 0 30 40 12 7 13 16 0 0 54 35
PR3	15	6 5 53 2 1 14 1 11 0 16 1 30 54 1 0 9 5 1 1 3 21 52 26 6 7 41 15 26
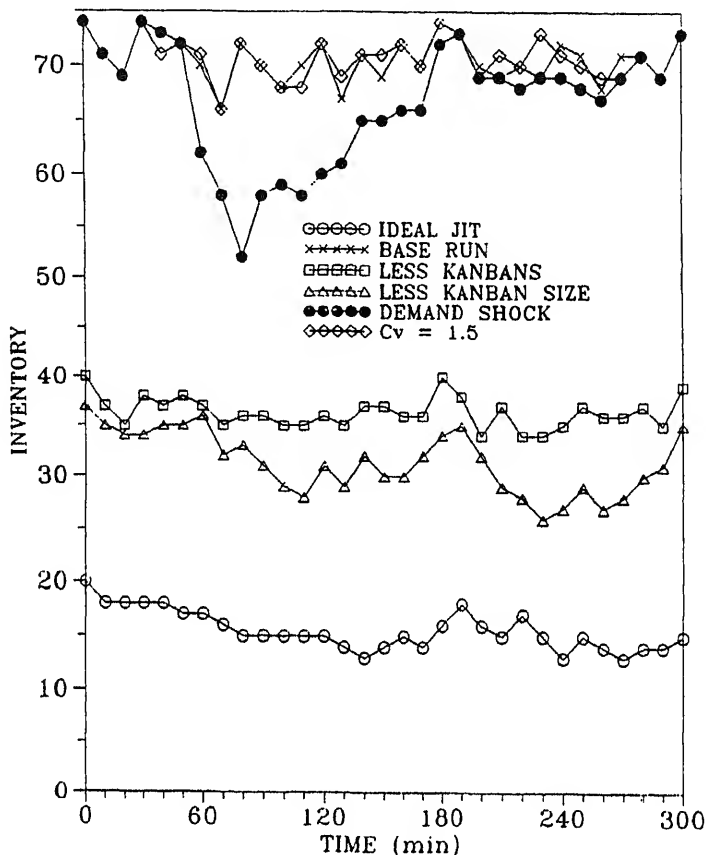
Figure 8. Variation of in-process inventory.

5. Analysis

The following design issues are addressed in the present study:

- Number of kanbans

- Kanban size

- Vehicle dispatching

Being a demand-driven system, the functioning of a JIT system starts with the arrival of a demand. A Poisson arrival pattern with a different mean arrival time for each part variety is considered in the present study. Inter-arrival times for each part type are shown in table 5. The same inter-arrival times are used in all simulation experiments in order to investigate the effects of different variables on system performance. In the initial analysis phase, several simulation runs are conducted to probe into the behaviour of the system.
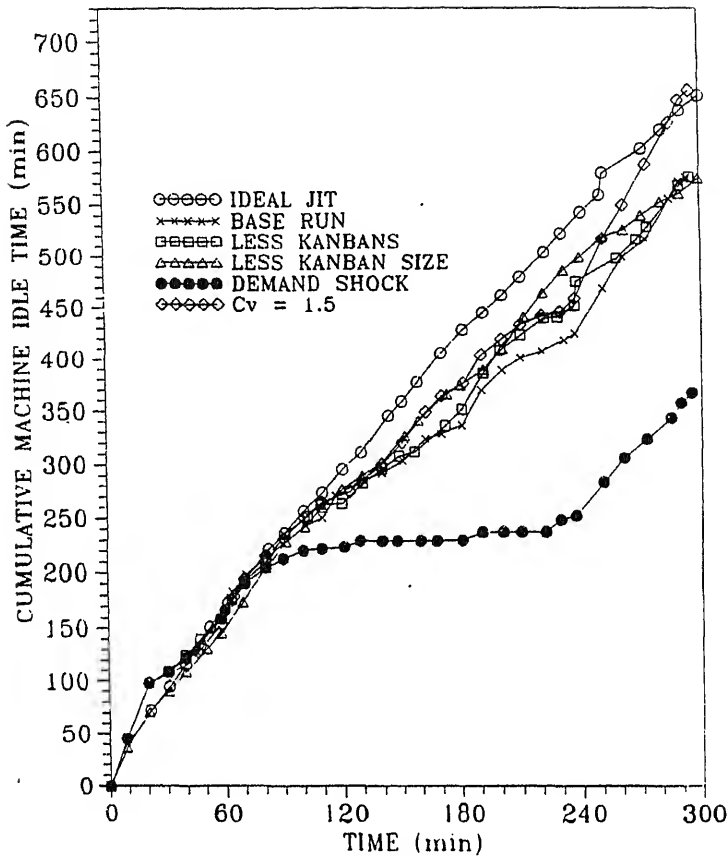
Figure 9. Variation of machine idle time.

(1) *Base run:* The purpose of the base run is to examine the model under typical conditions which will then serve as the basis for comparison in other experiments. The following parameter settings are used for this experiment.

(i) Number of AGVs = 3.

(ii) Number of kanbans = 2.

(iii) Kanban size = 2.

(iv) No variation in processing times.

In all the other runs that follow, the parameters are compared with those of the base run.

(2) *Run one (less kanban size):* The purpose of this run is to evaluate the effect of reduced lot size. The same settings are maintained as in base run except that the size of the kanban is reduced to one.

(3) *Run two (less kanbans):* In this experiment the effect of number of kanbans is investigated. With the original value of kanban size the number of kanbans is reduced to one.
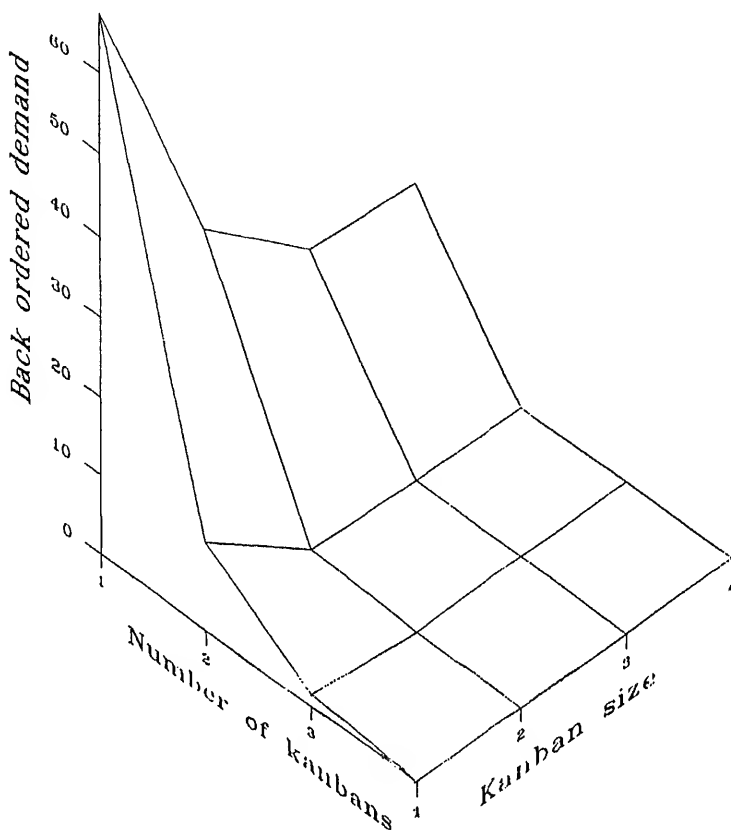
Figure 10. Variation of back ordered demand.

The intent of this test is to provide a contrast to the previous experiment (run one) and evaluate both the methods of reducing work in process inventory.

(4) *Run three (demand shock):* The purpose of this study is to evaluate the sensitivity of the system to demand shock. A demand shock (6 products of each part variety) is introduced at the beginning of the second hour. A JIT system, to be efficient, should support significant demand shocks for a short period of time.

(5) *Run four (Cv = 1.5):* The purpose of this run is to evaluate the effects of variability in processing times. All parameters of this run are same as that of the base run, except that there is 50% variability in processing times.

(6) *Run five (ideal JIT):* In this final experiment the system is evaluated with the smallest possible work-in-process level. This is accomplished with the following:
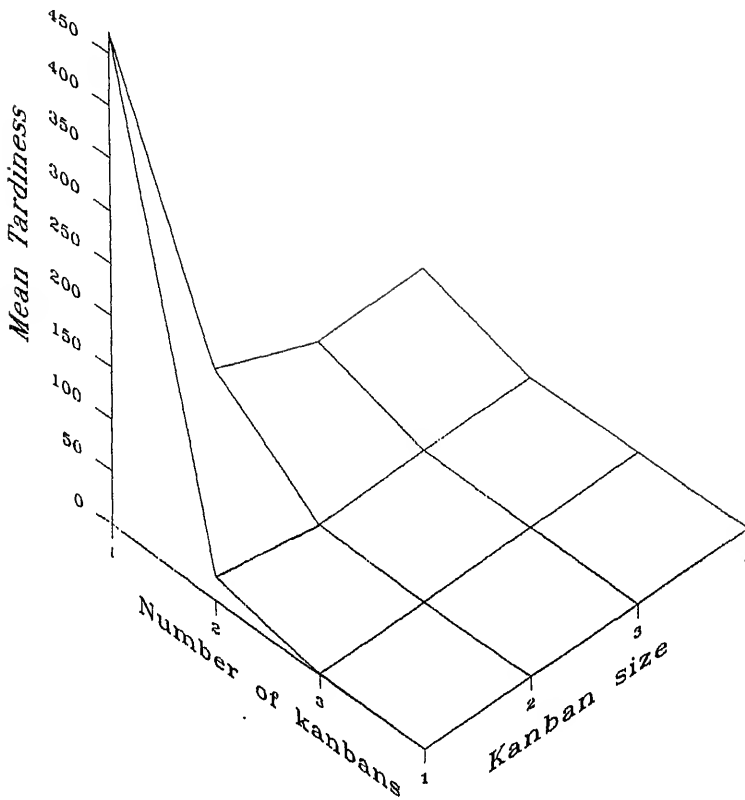
Figure 11. Variation of mean tardiness.

All the above simulation runs are compared with the help of two important performance measures namely in-process inventory and cumulative machine idle time. Figure 8 shows the comparison with respect to in-process inventory. Experiments with low kanbans and a kanban size of two (base run, run 3, run 4) have higher inventory levels. Runs with smaller kanban size and fewer kanbans (runs 1 and 2) have reduced inventory levels. Run 5 (ideal JIT) has minimum inventory and is as expected. An important feature of inventory can be observed in run 3, where there is a rapid decline of inventory during the second hour. This is due to the demand shock which comes into effect during this period and depletes the inventory. The comparison of runs 1 and 2 suggests that it is more effective to have a large number of kanbans with small size instead of a fewer kanbans with large capacity.

The other comparison is made with respect to cumulative machine idle time, figure 9. It is found to increase as the inventory reduces. The idle for the third experiment (demand shock) is minimum. Also, idle time remains unchanged during second and third hours. This is due to the fact that machines are kept busy during this period due to demand shock. It is found that idle time increases steeply in all other cases. This may be due to the fact that machines are kept busy during this period due to demand shock. It is found that idle time increases steeply in all other cases. This may be due to lower consumption rates and
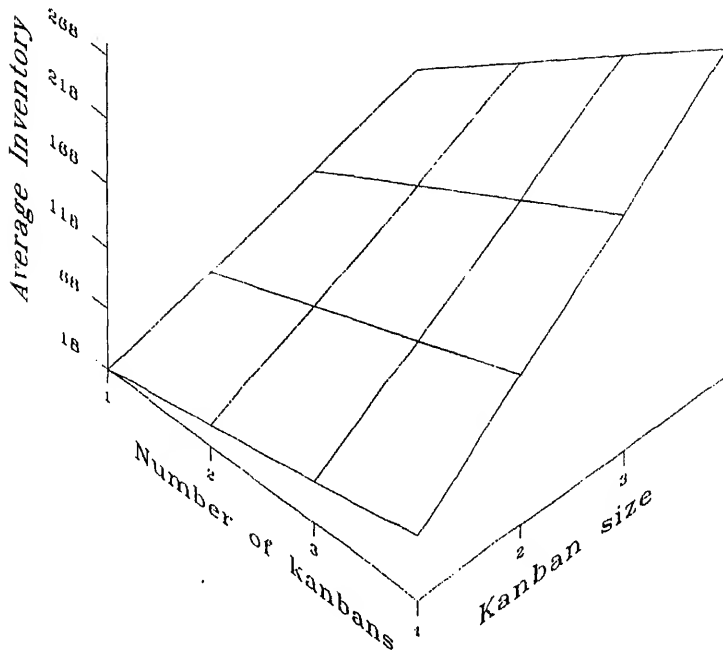
Figure 12. Variation of average inventory.

reduced number of AGVs. Another important finding is that the variability in processing times (run 4) results in increases in machine idle time.

Number of kanbans and kanban size: The three parameters considered for this analysis are percentage of back-ordered demand, mean tardiness and average in-process inventory. Three-dimensional plots in figures 10 to 12 show the effects of the number of kanbans and kanban size. As the number of kanbans increases, the percentage of back-ordered demand falls sharply at the beginning and reaches a minimum value. The same trend is observed with respect to kanban size.

Figure 11 shows the variation of mean tardiness. This varies in a similar fashion as in the case of back-ordered demand. Figure 12 shows that the in-process inventory increases sharply with increase in the number of kanbans and kanban size. Therefore, considering the in-process inventory, mean tardiness, mean back-ordered demand, two kanbans at every

Table 6. Effect of number of kanbans and kanban size.

Number of kanbans	Kanban size	Day's schedule			Completion time (s)	AMU*	AVU
		PR1	PR2	PR3			
2	1	54	37	36	23545	74.6	93.0
1	2	54	37	36	22953	76.2	51.9
2	1	40	40	40	21951	76.5	93.4
1	2	40	40	40	20724	83.2	57.4

AMU – Average Machine Utilization; AVU – Average Vehicle Utilization

Table 7. Effect of redundant work dispatching rules.

Redundant work dispatching rule	Cumulative machine idle time (min)	Back ordered demand (%)	Mean tardiness (s)	Average vehicle utilization (%)
ED	111.98	69.29	828.6	83.36
STT	105.98	66.92	463.3	80.15
MSPT	110.95	70.86	795.3	81.50
RJ	111.80	70.07	828.7	82.87

Job selection rule: SPT; Machine selection rule: LUM; Redundant vehicle dispatching rule: LUV

stage for each part variety, with size of kanban being '1', are found suitable for the system under study.

The analysis of the same system working to a daily schedule (single kanban) is also considered. In this case, the system acts as a push system for production and pull system for replenishments. Some typical results are shown in table 6. A kanban size of two with a single kanban at each stage is found to yield better results in this case.

Vehicle dispatching: The following dispatching rules are considered for the present study. Redundant work dispatching rules:

(1) Earliest Demand (ED);

(2) Shortest Travel Time (STT);

(3) Minimum work in queue (MWQ);

(4) Random Job (RJ).

Redundant Vehicle dispatching rules:

(1) Least Utilized Vehicle (LUV);

(2) Nearest Idle Vehicle (NIV);

(3) Longest Idle Vehicle (LIV); .

(4) Random Vehicle (RV).

Table 7 shows the effect of redundant work dispatching rules. STT is found to give better results with respect to vehicle utilization and other performance measures noted in this table. Table 8 shows the effect of redundant vehicle dispatching rules. LIV is found to perform well with respect to tardiness and back ordered demand. NIV is found to give better result with respect to minimum vehicle movement.

Table 8. Effect of redundant vehicle dispatching rules.

Redundant vehicle dispatching rule	Cumulative machine idle time (min)	Back-ordered demand (%)	Mean tardiness (s)	Average vehicle utilization (%)
LUV	105.98	66.92	463.3	80.15
NIV	106.79	67.17	470.3	78.35
LIV	107.17	64.57	354.8	79.49
RV	109.32	67.71	675.5	79.48

6. Conclusion

In this paper, an attempt is made to address the modelling, simulation and implementation issues of JIT in flexible manufacturing environment. Priority nets are proved to be suitable and competent tools for the modelling and analysis of kanban systems. Realizing the need for a thorough investigation of JIT in FMSs, an FMS implementing pull strategy is modelled. A large number of simulation runs are conducted to probe the behaviour of the system with respect to different parameter changes such as variability in processing times, demand shock, kanban size and number of kanbans. Implementation issues such as number of kanbans, kanban size, vehicle dispatching etc., are effectively addressed with the help of priority nets, thereby proving their utility.

References

Bitran G R, Chang L 1987 A mathematical programming approach to a deterministic kanban system. *Manage. Sci.* 33: 427–441

Drolet J, Montreuil B, Moodie R, Colin L 1990 Virtual cellular manufacturing layout planning. *Proc. of Industrial Engineering Conference*, San Francisco, California, pp 236–241

Drolet J, Montreuil B, Moodie C L 1991 Empirical investigation of virtual cellular manufacturing systems. *Proc. of International Conference on Computer Integrated Manufacturing*, Singapore, pp 323–326

Gentili E 1991 Flexibility and JIT: A new advantage for FMS. *Proc. of International Conference on Computer Integrated Manufacturing*, Singapore, pp 311–314

Gravel M, Price L W 1988 Using the kanban in a job shop environment. *Int. J. Prod. Res.* 26: 1105–1118

Gupta P Y, Gupta M 1989 A system dynamics model of a JIT-Kanban system. *Eng. Costs Prod. Econ.* 18: 117–130

Huang P Y, Rees L P, Tylor B W 1983 A simulation analysis of Japanese Just-In-Time technique (with kanbans) for a multiline, multistage production system. *Decision Sci.* 14: 326–344

Kim G C, Schniederjans M J 1990 A high performance programmable controller for CIM systems based on Petri net theory. *Proc. of IECON 89, 15th Annual Conference of IEEE Industrial Electronics Society*, Philadelphia, pp 805–810

Lulu M, Black J T 1987 Effect of process unreliability on integrated manufacturing/production systems. *J. Manuf. Syst.* 3: 189–211

Mascolo M D, Frein Y, Dallery Y, David R 1991 A unified modelling of Kanban systems using Petri nets. *Int. J. Flexible Manuf. Syst.* 3: 275–307

Masuyama A 1986 Idea and practice of flexible manufacturing system at Toyota. *Applying Just-In-Time* (ed.) Y Monden (Industrial Engineering and Management Press)

Mitra D, Mitrani I 1988 Analysis of a novel discipline for cell coordination in production lines. Technical report, AT&T Laboratories, NJ

Monden Y 1981 How Toyota shortened supply lot production time, waiting time, and conveyance time. *Ind. Eng.* 13: 22–29

Monden Y 1983 *Toyota production system* (Institute of Industrial Engineering: Industrial Engineering and Management Press)

Narendar K R, Mehra S, Frolick M N 1995 A comparative analysis and review of JIT 'implementation' research. *Int. J. Oper. Manage.* 15: 8–18

Ravi Raju K, Chetty O V K 1993 Priority nets for scheduling flexible manufacturing systems. *J. Manuf. Syst.* 12: 326–340

Schonberger R J 1982 *Japanese manufacturing techniques* (New York: Free Press)

Schroer B J, Black J T, Zhang S X 1985 Just-In-Time (JIT) with Kanban, manufacturing system simulation on a microcomputer. *Simulation* 45: 62–70

Sugimori Y, Kusunoki K, Cho F, Uchikawa S 1977 Toyota production system and kanban system materialization of Just-In-Time and Respect-for-Human system. *Int. J. Prod. Res.* 15: 553–564

Talavase J, Hannam R G 1988 *Flexible manufacturing systems in practice: Applications, design and simulation* (New York: Marcel Dekker)

Voss C A, Harrison A 1987 Strategies for implementing JIT. *Proc. of 4th European Conference on Automated Manufacturing*, Birmingham, pp 57–73

Wang H, Wang H P 1991 Optimum number of kanbans between two adjacent workstations in a JIT system. *Int. J. Prod. Econ.* 22: 179–188

White R E 1993 An empirical assessment of JIT in US manufacturers. *Prod. Inventory Manage. J.* 34: 38–42

Wilson I B 1986 Achieving flexibility in assembly through technology and people. *Proc. of 7th Int. Conference of assembly automation*, Zurich, pp 191–202

A re-entrant line model for software product testing

V V S SARMA[1] and D VIJAY RAO[2]

Department of Computer Science & Automation, Indian Institute of Science
Bangalore 560 012, India
[1]Present address: Tata Research Development and Design Centre (TCS), Plot
54B, Hadapsar Industrial Estate, Pune 411 040, India
[2]Present address: CASSA (DRDO), New Thippasandra PO, Bangalore 560 075,
India
e-mail: vvs@trishul.trddc.ernet.in; vvs@csa.iisc.ernet.in; vijay@cassa.ernet.in

Abstract. In today's competitive environment for software products, quality
is an important characteristic. The development of large-scale software products
is a complex and expensive process. Testing plays a very important role in
ensuring product quality. Improving the software development process leads to
improved product quality. We propose a queueing model based on re-entrant
lines to depict the process of software modules undergoing testing/debugging,
inspections and code reviews, verification and validation, and quality assurance
tests before being accepted for use. Using the re-entrant line model for software
testing, bounds on test times are obtained by considering the state transitions for
a general class of modules and solving a linear programming model. Scheduling
of software modules for tests at each process step yields the constraints for the
linear program. The methodology presented is applied to the development of a
software system and bounds on test times are obtained. These bounds are used
to allocate time for the testing phase of the project and to estimate the release
times of software.

Keywords. Software quality; software process modelling; re-entrant lines;
software product testing.

1. Introduction

In today's competitive environment for software products, quality has become an increasingly important concern to software development organizations. Quality denotes a multidimensional concept. As an intrinsic product attribute, the quality of software is recognized by the absence of defects. If we view quality from the point of product operation, attributes such as reliability, efficiency, usability and integrity are useful; whereas from the point of view of product transition/revision, parameters such as portability, reusability, inter-operability, and maintainability are important (Ghezzi *et al* 1988).

121

Several models relating to software quality have been proposed in the literature. These may be broadly classified into three categories, each for a separate purpose (Kan *et al* 1994).

(1) *Reliability models* for reliability assessment and prediction.

(2) *Quality management models* for managing quality during the development process. Quality management models are still in their development and maturing phase. These models emerged from the practical needs of large-scale development projects. The phase-based defect removal model and several tracking models belong to this category.

(3) *Complexity models* and metrics which are used by software engineers for quality assurance purposes. Complexity models explain quality from the internal structure and complexity of the software.

Software reliability modelling is.more mature than the other two types. A plethora of software reliability models have been developed over the years but, in spite of the extravagant claims for their efficacy, none can be trusted to give accurate results in all circumstances. An important reason for this is the validity of the assumptions underlying these models.

(1) A detected fault is immediately corrected.

(2) No new faults are introduced during the fault removal process.

(3) Reliability is a function of the number of remaining faults.

(4) Failure rate increases between failures.

(5) Testing is representative of the operational usage.

(6) Software is treated as a blackbox without looking at its structure and the process of its development.

Recently, there has been much emphasis on improving the software development process, with the assumption that this will lead to improved product quality. However, a precursor to improved processes is an understanding of the dynamics of current processes. With respect to software processes, there are two prevailing schools of thought (Bollinger & McGowan 1991):

- International Standards Organization (ISO) 9000 certification, and

- Software Engineering Institute (SEI) assessment based on the capability maturity model (CMM).

Process models for quality ensure the application of process engineering concepts, techniques, and practices to explicitly monitor, control, and improve the software process. However, these models do not yield quantitative measures of parameters such as reliability and usability to denote the quality of the product in the end.

Software development lifecycle is a model of the software process. There are many steps and activities in building a software product. The process followed to build, deliver and evolve the software product from the inception of an idea all the way to delivery and final retirement of the system is called the *software production process* and the order in which

these activities are performed defines the lifecycle for the product. Many models which attempt to capture this process, also called the *software lifecycle* models, have been developed. Such models are based on the recognition that software, like any other industrial product, has a lifecycle which extends from its initial conception to its retirement and that its lifecycle must be anticipated and controlled in order to achieve the desired qualities of the product. Dalal *et al* (1993) distinguish between the upstream phases comprising requirements, specifications and design, and downstream phases comprising coding, testing and maintenance of the software development process.

Conventionally, the software process is supposed to proceed sequentially from requirements to specifications, design, code, testing, and then to release. One extreme description of the process of software development conjures up the image of a *waterfall* flowing from requirements successively onto release with no feedback from a succeeding phase to a preceding phase. The other extreme envisions a *spiral* where feedback constantly loops back from a succeeding phase to a preceding phase as repair of the process is needed. In practice, the actual process could lie anywhere in between and one needs to accurately model the flow of software modules. This is analogous to the flow of silicon wafers undergoing processing (such as deposition, photolithography, etching etc.) in a semiconductor manufacturing plant. A study of software faults in the different phases of the lifecycle suggests that a majority of faults occur in the coding phase (Marick 1990) and that coding errors have substantially more severe effects than do design errors. Testing thus occupies a very crucial role in the overall software development process. The purpose of software testing is to detect errors in a program and, in the absence of errors, gain confidence in the correctness of the program. Efforts to improve the effectiveness of testing can yield substantial gains in software quality.

In this paper, we propose a queueing model based on re-entrant lines (figure 1) to depict the process of software modules undergoing testing/debugging, inspections and code reviews, verification and validation, and quality assurance tests before being accepted for use. This is the first model of its kind which depicts the *process* of testing software as seen in the software industry. The model takes into account the structure of the software, the individual modules being distinguished by their criticality in the mission and implementation, their usage in the operational field from profiles and test strategies used for testing these modules. We consider in our model, the notion of imperfect debugging and that new faults can be introduced in the process of imperfect debugging. The paper is organized as follows:
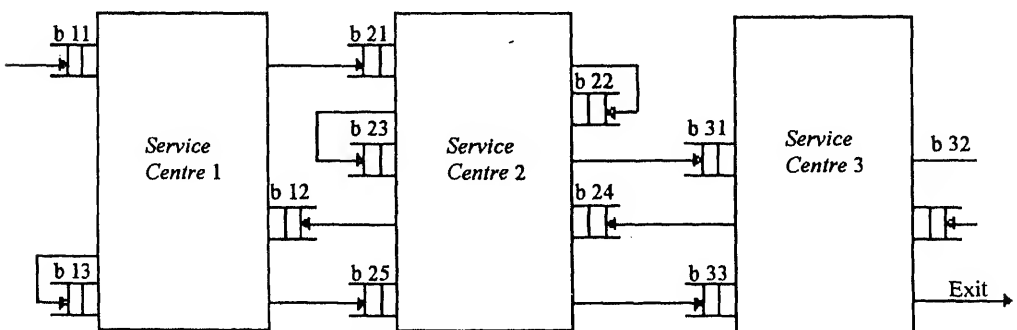
Figure 1. A typical re-entrant line.

Software development lifecycle is described in § 2, re-entrant lines and some important results are discussed in § 3, bounds on test times for software products using re-entrant lines are described in § 4, and a case study to illustrate the methodology is shown in § 5.

2. Software development process modelling

A large software project after requirement analysis and design is given to different programming teams for development. It is assumed that some software engineering methodology is used. The software is divided into modules based on the functions, its size and complexity. These modules after development need to be tested at various stages of the product building. Testing is done by the developers during the coding stage (local/unit testing). The module, after unit testing, is given to an independent test team, not involved in its development, for further testing. This independent test team detects the faults and these modules are sent back to the developers with a log of the tests done and their outcomes. This is also reflected in the configuration control management (CCM) of the project. The developing team then debugs the code and corrects the errors. The same sequence is followed for all the modules of the software. This process continues till the required reliability for the module is achieved or the testing time allotted for it is reached. Different criteria to stop testing have been suggested in the literature (Dalal & Mallows 1989; Musa & Ackerman 1989).

Once these modules are tested, they are integrated and tested for interface errors and inconsistencies across modules. These, along with the libraries and related documentation and standards, form the complete product. The validation of this product is done by an independent verification and validation (IVV) team. Code walkthroughs, inspections and quality assurance tests are done at all stages from coding to acceptance of the software product. These tests defer modules to further testing if they do not conform to requirements/standards prescribed, which would otherwise certify the product for release.

This whole process can be viewed as a multi-class queueing network as depicted in figure 2. The test teams denote the servers and the modules represent the customers who arrive for service (testing). In figure 2, the first team denotes the unit-testing team where the developers locally test the modules during its development, the second server represents the independent test team, the third team denotes integration tests and IVV; and the fourth team, the QA and system testing.

Consider the flow of a tagged module M through such a process. At the first test team, TT1, the module is unit tested by the developers. This module M is tested by an independent test team TT2 and the errors (if any detected and located) corrected by TT1. Unit tested modules arrive for integration and later for verification and validation. Interface errors, non-conformance with requirements, or inconsistent representational formats with some modules causing integration tests to fail result in these modules being sent back to the corresponding teams for correction. Finally, when all the modules are integrated, the system
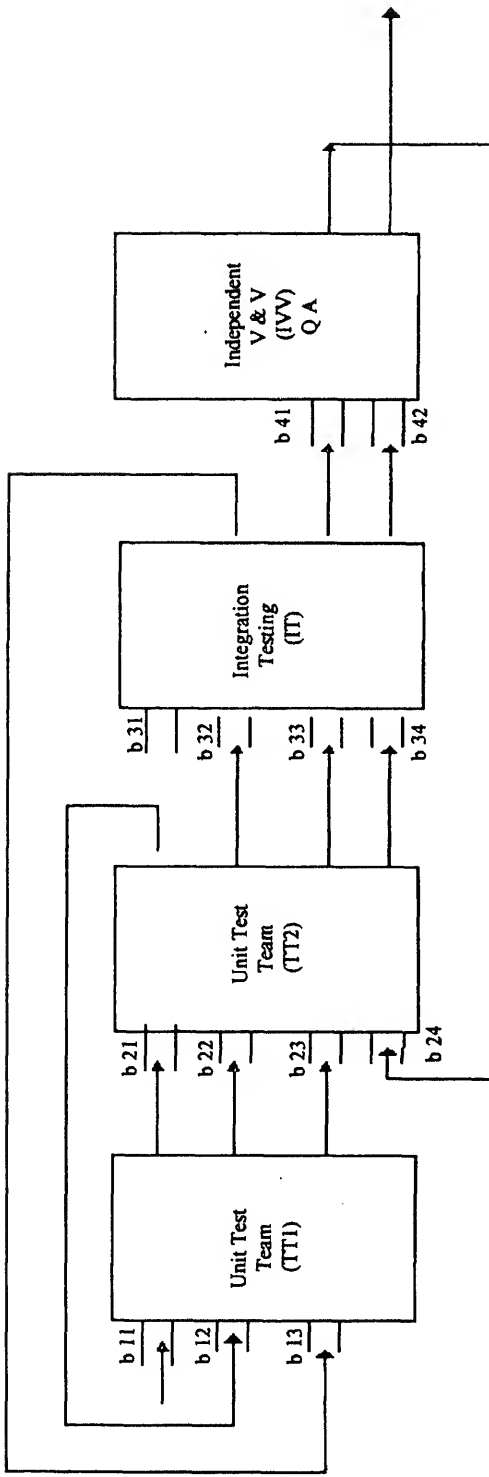
Figure 2. A re-entrant line model for downstream lifecycle phases for a software product with several modules.

3.   Model development

Semiconductor wafer manufacturing plants are organized quite differently from traditional assembly lines or job shops. The production process of a silicon wafer consists of imprinting several layers of chemical patterns on the wafer; the final end product obtained is a multi-layered sandwich. Each layer in turn requires several steps of individual processing such as deposition, photolithography, etching, etc. with many of the steps repeated at several of the layers. The machines to perform these individual steps are very expensive. Hence, the machines are not replicated but revisited by the wafers for processing at different layers. The distinguishing characteristic of such a manufacturing system (modelled as a multi-class queueing networks), called a *re-entrant line*, is that the lots revisit several machines at several stages of their life. The main consequence of the re-entrant nature is that several wafers at different stages of their life have to compete with each other for the same machines. Figure 1 shows a re-entrant line with 3 service centres and 11 buffers. Parts enter the system at buffer b_{11} and visit the centres according to a deterministic route as shown. Finished parts emerge from centre 3 after undergoing processing following a wait in b_{33}. Note that each part in this example line visits centre 1 three times, centre 2 five times, and centre 3 thrice. Scheduling in re-entrant lines, input releases and scheduling policies have a significant effect on the performance of this system. Several policies have been studied by Kumar (1994), Lu *et al* (1991) and Khan (1995).

Several researchers have recently come up with analytical methods to obtain upper and lower bounds on the performance of scheduling policies in multi-class Markovian networks (re-entrant lines) (Kumar & Kumar 1994). These methods rely on assuming stability and obtaining a set of linear constraints on the mean values of certain random variables that determine the performance of the system. Augmenting these constraints with others obtained using conservation principles, bounds on performance can be obtained by solving the resulting linear program. Bounds on the mean delay (called cycle time) are obtained with different scheduling policies. The cycle time in software testing process corresponds to the time required to test all the software modules before release to the customer.

In the proposed model for software product testing, servers (machines) denote the test teams and parts (silicon wafers) denote the software modules undergoing testing and correction. Due to the large number of modules at different stages of testing, test teams also need to schedule their tasks to select the next module to test.

Consider a set of $\{1, 2, 3, \ldots, S\}$ of S test teams consisting of professionals and developers of the code. Modules are classified based on their criticality and usage (from profiles). Modules of similar reliability requirements enter the system for testing at a test centre $s(1) \in \{1, 2, 3, \ldots, S\}$ where they are labelled as of class type C_i. Let C_L class of modules being tested at $s(L)$ be the last set of tests done on these modules. The sequence $\{s(1), s(2), \ldots, s(L)\}$ is the route followed by the modules for tests. These modules visit the next team $s(2)$ after being tested at $s(1)$ and so on. We shall allow for the possibility that $s(i) = s(j)$ for some classes $i \neq j$ and accordingly call this type of system a re-entrant line (Kumar 1993).

For this system we assume:

(1) Modules arrive into the system for testing according to a Poisson process with rate λ;

(2) The mean time to test for every class C_i is $1/\mu_i$ and the times to test are distributed exponentially.

The first team denotes unit/local testing which is done by the programmer himself/ herself. These modules take $1/\mu_1$ amount of time for the local testing. After this testing, the module is passed on to Team 2 which is a peer test team, not involved in the development of the module. Any bugs located by this team are recorded in the error log and sent back to the developers for correction and testing thereof. The time for testing this module would now be governed by the mean time to test for modules of class C_2. In this fashion, when the modules are approved by the Teams 1 and 2, it passes on to Team 3 denoting Integration testing and System testing. Team 4 denotes Product testing and QA which checks for the process of software development and the product developed. This team either accepts the product in which case it is delivered to the user along with the proper documentation, or it sends back particular parts of the product which have non-conformance reports to the design team or for further testing. This feedback defines the re-entrant path for the module. This completes one cycle of the downstream process for the software. Due to non-conformance of some modules to the specifications/standards, the product release date is shifted till another cycle of the process is completed. However, in Cycle 2, the mean time to test for some classes is less than that of Cycle 1, due to the learning factor, experience gained and familiarity with the system to generate efficient test cases which maximize the coverage. This is analogous to the product-in-a-process approach suggested by Laprie (1993) to develop families of software. The path followed by modules demanding different levels of quality in this process is varied. Based on this model, we compute bounds on mean test time for modules.

4. Bounds on test times: The LP approach

Consider a strategy to select the next module for testing, which is –

(1) *Nonidling*: If there is any module to be tested then the test team does not stay idle;
(2) *Stationary*: The decisions to select the next module depends only on the number of modules of different classes in the system (Lu *et al* 1991; Kumar 1993, 1994).

Let us rescale time so that $\lambda + \sum_{i=1}^{L} \mu_i = 1$. We use uniformisation in which we sample a continuous time system to obtain a discrete time system with the same steady-state behaviour. We sample the system at all service completion times, as well as at the arrival times of new modules to the system for testing. Let $\{\tau_n\}$ be the sequence of such random sampling times and let F_{τ_n} denote the σ-field generated by the events up to time τ_n. Let $X_i(t)$ denote the number of modules of class C_i at time t. Also, let $W_i(\tau_n) = 1$ if the testing team at $\sigma(i)$ is working on the module of class C_i at time t, and 0 otherwise. We take all processes to be right continuous, and thus $X_i(\tau)$ is the state after the nth event, while due to the stationarity of the strategy chosen, $W_i(\tau_n) = 1$ implies that the team $\sigma(i)$ is busy working on C_i class of modules in the interval $[\tau_n, \tau_{n+1})$. Let us denote

$$X^T(\tau_n) = (X_1(\tau_n),\ X_2(\tau_n), \ldots, X_L(\tau_n)). \tag{1}$$
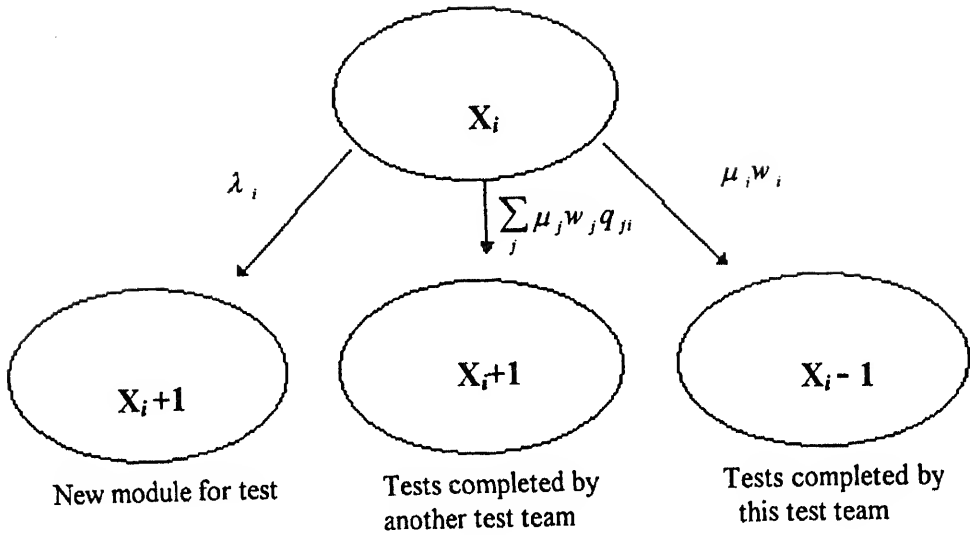
Figure 3. State transitions for a class C_i of modules.

In the steady state,

$$E[X^T(\tau_{n+1}).Q.X(\tau_{n+1})] = E[X^T(\tau_n).Q.X(\tau_n)],$$

for every symmetric matrix Q. We presume that the steady-state distribution has a fir second moment on the total number of modules at each buffer. For this equation to ho we need

$$E[X_i(\tau_{n+1}).X_j(\tau_{n+1})] = E[X_i(\tau_n).X_j(\tau_n)] \quad \text{for } 1 \le i, \ j \le L.$$

Now consider the implication of the equality

$$E[X_1^2(\tau_{n+1})] = E[X_1^2(\tau_n)].$$

From the state transitions for the buffer i shown in figure 3, we have

$$
\begin{aligned}
X_i(\tau_{n+1}) &= X_i(\tau_n) + 1 \text{: exogeneous arrival to } C_i \text{ at } \tau_{n+1}, \\
&= X_i(\tau_n) + 1 \text{: previous class tests completed}, \\
&= X_i(\tau_n) - 1 \text{: current class tests completed}, \\
&= X_i(\tau_n) \quad \text{: otherwise.}
\end{aligned}
$$

Suppose every class C_i has an exogenous arrival process, which is Poisson with rate Also suppose that with probability q_{ij}, a module passes from class C_i to C_j.

From the equality equation, $E[X_1^2(\tau_{n+1})] = E[X_1^2(\tau_n)]$, and using the stationa policy of the strategy used, we get the following equality constraints:

where $z_{ij} = E[W_i(\tau_n).X_j(\tau_n)]$

$$\lambda_i \left(\sum_{k \in I(j)} z_{kj} \right) + \lambda_j \left(\sum_{k \in I(i)} z_{ki} \right) + \sum_{k \neq j} \mu_k q_{ki} z_{kj} + \sum_{k \neq i} \mu_k q_{kj} z_{ki}$$
$$+ \mu_j q_{ji} (z_{jj} - z_{ji} - \rho_j) + \mu_i q_{ij} (z_{ii} - z_{ij} - \rho_i)$$
$$- \mu_i (1 - q_{ij}) z_{ij} - \mu_j (1 - q_{ji}) z_{ji} = 0.$$

Now using the nonidling policy, we get the following inequality constraints:

$$\sum_{\{j | \sigma(j) = \sigma\}} z_{ji} \leq \sum_{j \in I(i)} z_{ji}, \quad \text{for } i = 1, \ldots, L; \quad \sigma = 1, \ldots, S \text{ with } \sigma \neq \sigma(i),$$

$$(5)$$

and the nonnegativity constraints

$$z_{ij} \geq 0 \quad \text{for } i, j = 1, \ldots, L. \tag{6}$$

If the scheduling strategy is stationary and nonidling with a steady-state distribution possessing a finite second moment, then the mean number of modules in the system at various stages of testing is bounded above by

$$\max \sum_i \sum_{j \in \sigma(i)} z_{ji}, \tag{7}$$

and below by

$$\min \sum_i \sum_{j \in \sigma(i)} z_{ji}. \tag{8}$$

Equations (7) and (8) denote the bounds on the number of modules in the system. Using Little's law, $L = \lambda W$ (Little 1961) and assuming that the arrival rate of modules to test is constant, we obtain the bounds on testing time for the modules.

5. Examples

Example 1. In this section, we consider the development of a re-entrant line based software process model for a firm executing a software project of moderate size (needing a few person months of effort). It is identified at the preliminary design level that the software is made up of 40 modules of similar complexity. The underlying re-entrant line model is shown in figure 4. It is assumed that there are two programming and testing teams. Software modules are first unit tested by the developers (Team 1) and Team 2 acts as an independent test team for these modules.

For simplicity, we assume that new modules arrive for testing by Team 1 with rate λ and the route followed by all modules in the re-entrant line model is deterministic. A class C_i of modules takes $(1/\mu_i)$ person hours to test a module. The linear program to bound the mean number of modules in the re-entrant line (figure 4) is:

$$\min[z_{11} + z_{31} + z_{22} + z_{42} + z_{13} + z_{33} + z_{24} + z_{44}]$$
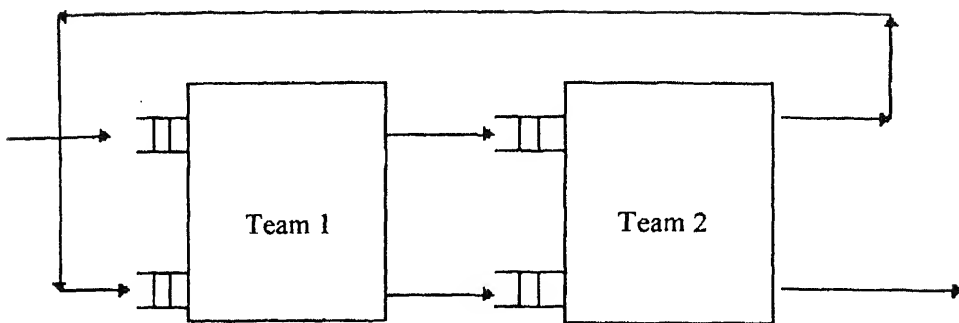
Figure 4. A re-entrant line model for a software testing process.

and

$$\max[z_{11} + z_{31} + z_{22} + z_{42} + z_{13} + z_{33} + z_{24} + z_{44}].$$

The equality, inequality and non-negativity constraints are given by (4), (5) and (6) respectively. Defining $\rho_i = \lambda/\mu_i = 0.25$ and solving the linear program, we obtain bounds test times as [4.0–4.5] person months. With these estimates of test times, we can alloc approximately 18 person weeks for the testing phase of this project.

Example 2. If the 40 modules in the software system of example 1 are classified bas on their criticality and usage, with more test time allocated to the high usage and criti modules, we can obtain realistic values for bounds on test times. We classify the modu based on the criticality of their function in the mission and the usage (from profiles) in use environment. The *Criticality-Usage* matrix is formed for the modules of the syste This was used to decide the *class* to which the module enters and hence the testing tir These data are summarized in table 1.

In the above matrix, modules of $\{CU(1, 1)\}$ are made members of class C_1, m ules of $\{CU(1, 2),\ CU(2, 1),\ CU(2, 2)\}$ are made members of class C_2 and modules $\{CU(1, 3),\ CU(2, 3),\ C(3, 1),\ CU(3, 2),\ CU(3, 3)\}$ are made members of C_3, as th are critical and frequently used modules. The testing times are varied accordingly in ratio of 1:2:4 for the modules of classes $C_1: C_2: C_3$. The linear program is solved to obt the bounds on the test times for the modules of different criticality and usage. The resu obtained are summarized in table 2.

Table 1. The criticality-usage matrix $CU(i, j)$.

	Criticality		
Usage	Low	Medium	High
Low	6	10	2
Medium	4	6	2
High	4	3	3

Table 2. Bounds on test times for example 2 (§ 5).

Class	Mean test time in person weeks (bounds)
C_1	[3, 4]
C_2	[8, 11]
C_3	[19, 23]

Total test time: [30, 38] person weeks or [7.5–9.5] person months (Assume 4 person weeks in a person month)

ith these estimates of test times, we can allocate approximately 9.5 person months of ng. If the milestone for completion of coding is set at the end of 14th month, then we allocate the testing and verification phase to end by the 24th month from the start of project.

Conclusions and discussion

ocess model which depicts the downstream phases of the software life cycle modelled re-entrant line is presented. Further, based on this model, a method to compute bounds est times of software is presented. Due to priority test scheduling of modules, the re-ant model is not of product form and hence not amenable to closed form solutions for dy-state analysis. Bounds on test times are obtained by considering the state transitions general class of modules which leads to a linear programming model. Scheduling of ware modules for test at each process step yields the constraints for the linear program. n the bounds on the test times, the product release times are obtained. We illustrate methodology using an application for which bounds on test times are obtained. For lules of varying criticality-usage factor, we observe that the *test times are not scaleable*. software development applications, a module's route through test teams is not the e for all modules and is not deterministic. The current model can be extended to reflect situation with the introduction of path profiles and a route matrix for the modules ay Rao 1995). This model can also be used to decide on the release times of software a specified reliability measure (Vijay Rao 1995).

erences

inger T B, McGowan C 1991 A critical look at software capability evaluations. *IEEE Software* : 25–41

l S R, Mallows C L 1989 When should one stop testing? *J. Am. Stat. Assoc.* 83: 872–875

l S R, Horgan J R, Kettenring J R 1993 Reliable software and communication: Software uality, reliability and safety. *Proc. 15th Int. Conf. Software Engineering* (Los Alamitos, CA:

Ghezzi C, Morzenti A, Pezze M 1988 On the role of software reliability in software engineering. *Software reliability modelling and identification* (ed.) S Bittanti (Berlin: Springer-Verlag) pp 1–41

Kan S H, Basili V R, Shapiro L N 1994 Software quality: An overview from the perspective of total quality management. *IBM Syst. J.* 33: 4–18

Khan L M 1995 *Performance analysis of scheduling policies in stochastic re-entrant lines.* Ph D dissertation, Indian Institute of Science, Bangalore

Kumar P R 1993 Re-entrant lines. *Queueing Syst. Theor. Appl.* 13: 87–110

Kumar P R 1994 Scheduling queueing networks: stability performance analysis and design. *Proc. IMA workshop on stochastic networks* (Berlin: Springer-Verlag)

Kumar S, Kumar P R 1994 Performance bounds for scheduling queueing networks. *IEEE Trans. Autom. Control* 39: 1600–1611

Laprie J C 1993 For a product-in-a-process approach to software reliability evaluation. PDCS Tech. Report ESPRIT-BRA-6362-PDCS2, University of Newcastle upon Tyne

Little J D C 1961 A proof of the queueing formula $L = \lambda W$. *Oper. Res.* 9: 383–387

Lu S C H, Ramaswamy D, Kumar P R 1991 Efficient scheduling policies to reduce mean and variance of cycle-times in semiconductor manufacturing plants. *IEEE Trans. Semiconductor Manuf.* 7: 374–388

Marick B 1990 A survey of software faults. Report No. UIUCDCS-R-90-1651, Dept. of Computer Science, Univ. of Illinois, Urbana, Champaign

Musa J D, Ackerman A F 1989 Quantifying software validation: When to stop testing? *IEEE Software* 5: 19–27

Vijay Rao D 1995 *Estimation of software release times based on a queueing model for software testing.* MSc (Eng.) thesis, Indian Institute of Science, Bangalore

Competitive manufacturing systems – Part 2

Foreword

This is the second of the two special issues in the important area of competitive manufacturing systems. The present issue contains eight papers covering several important topics. The first two papers view a manufacturing enterprise as a collection of value-delivering business processes and propose quantitative methods for designing world-class enterprises. The next three papers focus on new technologies for computer-aided design and engineering of high quality products. The following two papers are concerned with efficient algorithms for machine and job shop scheduling. The eighth paper in this special issue looks at recent advances in material handling systems, which constitute an essential subsystem in manufacturing plants.

In the first paper of this issue, Viswanadham and Raghavan emphasize the important notion of flexibility. There has been a fair amount of research on flexibility of manufacturing systems in recent times, but the research has looked at only subsystems in an enterprise. The work of Viswanadham and Raghavan takes a process-based look at flexibility and proposes qualitative and quantitative models to assess and manage flexibility of business processes. Three critical business processes are discussed: New product development process, order-to-delivery process, and supply chain process. After defining generic performance measures, including flexibility, for business processes, the authors identify the various types of flexibilities by emphasizing these three business processes. The authors also advocate strategies for managing flexibility based on the insights revealed by the quantitative models.

In the second paper, authors Bhaskaran and Leung concentrate on the supply chain process and present quantitative models to facilitate reengineering of supply chains. The models that the authors employ are derived from the classical operations research/industrial engineering disciplines. Through several examples and a case study, the authors demonstrate the use of these models in identifying reengineering opportunities, evaluating design alternatives, guiding the selection of the best alternative, and deploying tools to implement the design.

The third paper in the special issue is by Hock. In this paper, the author argues and demonstrates that design, engineering, and manufacturing are all closely related. By concentrating on two well-known design methods, namely quality function deployment and Pugh's concept and using a case study of designing and developing a hematology machine, the author makes a case for integrated product development.

In the next article of this issue, Chandru and Manohar provide an overview of some major and some recently emerged paradigms for geometric modelling: Constructive solid

geometry, boundary representation, non-manifold models, and voxel models. The authors then elaborate on the unique virtues of voxel models in the context of emerging manufacturing technologies and argue strongly in favour of voxel models. The authors also look at important research directions in voxel modelling.

Narayan, Rao, and Gurumoorthy, in their paper on feature-based geometric reasoning for process planning, present a framework based on domain-independent form features for automatic assessment of manufacturability and process planning for machining. Their approach offers a novel improvement over existing approaches by combining automatic extraction of domain-independent features with knowledge-based reasoning. This facilitates an effective framework for automated process planning.

The next two papers are on the topic of scheduling. The first paper, by Wang, Luh, Zhao, and Wang, originates a Lagrangian relaxation based scheduling methodology for job shop scheduling by considering on-time delivery and low work-in-process as the objectives. The methodology is based on iterative solving of decomposed subproblems and the authors present an algorithm that combines backward dynamic programming with the conjugate gradient method to solve the scheduling problem in a Lagrangian framework. The method has very good performance in terms of rapid convergence and automatic evaluation of the quality of schedules obtained. Their methodology is a rich addition to the collection of techniques available for this difficult problem.

In their paper on the on-line maintenance of optimal machine schedules, authors Aman, Balakrishnan, and Chandru consider the on-line scheduling problem of dynamically scheduled jobs with precedence constraints on a single machine to minimize the maximum completion time. They develop an efficient technique to reoptimize a schedule when new jobs arrive. The paper illustrates many algorithmic and analytical issues associated with developing effective methods to update schedules on-line.

Material handling is a critical component in modern manufacturing plants because it provides material integration in the factory and can help cut lead times and inventory. There have been significant advances in the technology and operation of discrete material handling systems in recent times. Heragu and Rajagopalan provide a very useful overview of the advances in this area, in the eighth paper of this special issue.

It has been a pleasure for us to work with all the individuals concerned with various aspects of this two-volume special issue. First, we would like to acknowledge the continuous support and ideas from Professor Viswanadham. Authors of papers have been most co-operative and have taken minimal time. Reviewers have been very prompt and thorough in their reviews. It is mostly because of this wonderful quick response from the authors and reviewers that these special issues have seen the light in exactly one year's time (from conception to market). The Academy staff, coordinated efficiently by Ms Shashikala, have also contributed significantly to the compression of this issue's lead time. We wish to place on record the excellent cooperation from our respective departments.

April 1997

Y NARAHARI
M KAMATH
RAVI ANUPINDI
Guest Editors

Flexibility in manufacturing enterprises

N VISWANADHAM and N R SRINIVASA RAGHAVAN

Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India
e-mail: [vishu,raghavan]@csa.iisc.ernet.in

Abstract. A manufacturing enterprise is a collection of interrelated, flexible, optimized business processes delivering value to the customers through high quality products and services, faster than competition. This view of an enterprise enables one to consider the entire business system including the suppliers, product development, manufacturing, logistics, distribution, and retailing and to smoothen out the interfaces between them. Performance measures and performance measurement are important for monitoring, control and management. We identify and discuss eight performance measures for generic business processes. These include lead time, customer service, dependability, quality, flexibility, cost, capacity, and asset utilization.

In this paper, we concentrate on flexibility of business processes with special emphasis on the supply chain and order-to-delivery processes. We attempt to provide clear definitions and measures of various types of flexibilities as well as discuss the relationship between product structure and supply chain flexibility. The relationship between uncertainties, flexibility, technology, and product structure is clearly brought out in this paper.

Keywords. Manufacturing enterprise; performance measures; flexibility.

1. Introduction

It is commonly believed that there will be dramatic changes in the ways successful manufacturing companies are organized in the 21st century. These changes are basically driven by technological innovations, changes in political landscape, and demographics. In the developed world, electronic commerce has proliferated, and the information superhighway has made access to video connection and text-based information easy and almost free. Organizational structures, accounting practices, trading methods, compensation and incentive schemes that were so successful in the decades after World War II have become obsolete. In an earlier report (Viswanadham 1996), we have delineated issues that make manufacturing play a proactive role in inventing new businesses and organizational structures that give sustainable competitive advantage.

Manufacturing has gone through successive periods of great changes. New materials, such as plastics, ceramics and composites, new technologies such as computer-aided design, manufacture and inspection, and the internet, new techniques such as kanban and just-in-time, new bases for competition such as cost, quality, time or core competence have all been at the root of these changes. Currently, global competition, demanding customers, liberalization which provides a congenial environment for direct foreign investment, regulations on environment, emergence of common markets, disintegration of large states, and volatile exchange markets have made manufacturing a more complex function. Customers want it all: low cost, low defect rates, high performance, on-the-spot delivery and maintenance without irritants. To meet such demanding customer needs, computer-aided automation, effective flexibility management, strategic alliances, management of end-to-end processes such as supply-chain process, new product development process etc., are important.

This paper is organized as follows: In § 2, we survey the history of manufacturing emphasizing the paradigm shift from mass production to mass customization. In § 3, we describe the cosmic view of a manufacturing system including suppliers, distributors, employees, technology, competition and customers. In § 4, we define a manufacturing system as a collection of business processes and consider issues of analysis using this approach.

In § 5, we describe the supply-chain process and the order-to-delivery process which are two important customer value delivery processes in a manufacturing company. In § 6, we identify eight performance measures for a business process. Section 7 is on flexibility of business processes. Since flexibility is the effective management of change, it is essential to conduct root cause analysis of changes and find ways of neutralizing or coping with the changes. We define business process flexibility in general and discuss the influence of technology on flexibility. We bring out the importance of information-sharing and communication in enhancing the flexibility of business processes. In § 8, we present the definitions of various flexibilities of the supply-chain and the order-to-delivery processes. Product structure is the subject matter of § 9 illustrating the influence of product modularity on staged manufacturing and the supply chain. We conclude the paper in § 10.

2. History of manufacturing

Manufacturing, the world over, has undergone tremendous changes. Several companies in the West have gone through the phases of high growth, decline, restructuring, steady growth and so on. Several revolutions have occurred in the manufacturing arena: just-in-time, total quality management, time compression etc. In this section, we survey the history of manufacturing with emphasis on changes in factors driving the competition and changes in organization structure.

2.1 *Mass production system*

In the early twentieth century, Henry Ford revolutionized manufacturing with the introduction of the transfer line for mass production. Complete and consistent interchangeability

of parts and their assembly on a moving, continuous assembly line gave Ford tremendous advantages over competitors. Specialization and division of labour were the key concepts based on which both blue and white collar jobs were organized. Unskilled labour were trained quickly and supervisors, industrial engineers and quality inspectors ensured consistency and accuracy. The white collar jobs (so-called back-office work), were also broken down into small, repeatable tasks which were mechanized and automated. F W Taylor invented the principle of scientific management which made work more specialized, precisely defined, interchangeable and optimized. Alfred Sloan as the President of General Motors was responsible for developing the easily scalable, pyramidal organizational structure by creating decentralized divisions managed by specialists and coordinated by corporate headquarters. The mass production system was thus born and the fathers were Ford, Sloan, and Taylor (Womack *et al* 1990).

The mass production system is characterized by assumptions of constancy and predictability of demand, and the logic of economies of scale and division of labour. The characteristics of such factories include: dedicated machines, long production runs, narrow product range, low skilled workers, command and control management, vertical internal communication, high volume, sequential product development, high inventories, make to stock, limited communication with the customer, large number of suppliers and dealers, mass marketing etc. This mass production paradigm has influenced generations of industrialists and has generated wealth for western nations for over four decades. Also, we can see that stable, homogeneous markets and standardized products with long life, dedicated mass production facilities, command and control organization structure, low skilled labour, and long production runs, are all mutually reinforcing (Milgrom & Roberts 1990).

2.2 *Modern manufacturing system*

The face of manufacturing has changed. Dedicated equipment is replaced by flexible machine tools and programmable multitasking production equipment (Viswanadham & Narahari 1992). Their use reduced changeover times and small batch size production became economical. Small batch sizes shortened production cycles and reduced work-in-process and finished goods inventories. Advances in computers and communications made possible direct contact with customers, suppliers and dealers. Rapid transport and communications resulted in global competition. This has created customers who make relentless demands in quality, service and price contrasting with the gentle, grateful, loyal customers of Ford and Sloan. Toyota, under the leadership of Ohno and Toyoda perfected the *lean* manufacturing system with emphasis on just-in-time deliveries, quality, and planning production and product development jointly with suppliers and dealers. Manufacturers now use the point of sale information to determine the production schedules. There is a general strategic emphasis on speeding up all aspects of a firm's operations: Shorter development cycles, quicker order processing, freedom from defects and speedier delivery. Make-to-order and almost instantaneous delivery have become more the rule than the exception. Frequent product improvements and new product introductions, combined with the need for speed have resulted in the use of cross functional teams for design and manufacturing.

We see thus that Adam Smith's world has changed considerably. Companies created to thrive on the mass production paradigm cannot succeed in the world of fast changing

customer demands, short product life cycles, changing technologies, fierce competition and fluctuating exchange rates.

2.3   *Competitions based on cost, quality, time and customer focus*

Lean production or Toyota production system has introduced the concept of waste elimination by using just-in-time production. Waste, defined as any non-value adding activity, such as storage, inventory, transport, inspection, setups, machine down time, repair etc. should either be eliminated or minimized. Toyota has relentlessly attacked all forms of waste; reduced setup times, given importance to '*doing it right first time*', eliminated inventories, introduced kanbans to trigger production, and used load leveling methods to smoothen the work flow. Toyota's methods of cost reduction during the energy crisis years of the 70's made them leaders of world class manufacturing. Also the new concept of target costing i.e. price is not cost plus profit but cost is market price minus profit, has been introduced (Womack *et al* 1990).

The eighties have seen the total quality management (TQM) revolution spreading from Japan to the West. TQM strives for a totally integrated effort towards continuous improvement at every level of the company: Design, production, marketing and sales. The ultimate beneficiary of TQM activity is the customer who receives high quality products/service at a reasonable cost. TQM involves process control rather than product testing. The aim is to produce zero-defect products through statistical process control and coordinated testing at input, production and final delivery points. Company-wide education and training, management commitment to design and implementation of an effective quality, performance measurement and reward system are essential for the success of TQM programs (Smith 1993).

Flexible automation involving numerically controlled machines connected by an automated material handling system and local area networks was the next important development. Computer-aided design and manufacture has made possible production of a variety of products at mass production economies. Consistent quality and manufacturing flexibility are two distinguishing features of factory automation. Together with factory automation, production planning and control systems such as MRP-II also emerged. The MRP-II system includes bill of materials, production plans and schedules, shop-floor control, inventory analysis, forecasting, purchasing, order processing, cost accounting, capacity planning etc. at various levels of detail (Viswanadham & Narahari 1992).

Time-based competition is the issue occupying the minds of manufacturing executives even today. Customers want it all: price, quality and faster delivery, and hence firms must shrink the time from conception to consumption. Time reduction provides an important leverage not available in cost-reduction strategies. By removing time from their operations, costs are reduced. In manufacturing industries, time compression requires that attention be given to all activities including order processing, scheduling, distribution and customer service, i.e. time compression should include all activities not just on a single production

A new way of organizing businesses based on core competences is emerging in contrast to the product-centric view currently held by most companies (Prahlad & Hamel 1990). Here, businesses nurture a bundle of competences by developing skills, capabilities, which in turn allow the company to market a group of world class products. Core competences developed over time by integrating the skills and resources in an organization will give the company a sustainable competitive advantage. Honda's knowhow in engines; logistics and inventory management of Federal Express; Sharp's competence in flat-screen displays etc., have allowed these companies to manufacture a variety of products/services which appear diverse. Many companies are turning to the competence-centric view and are disinvesting in the non-core sectors of their businesses.

We thus see that the logic of modern manufacturing is rooted in time and competence-based competition, flexibility and scope economies. Over the last decade a paradigm of mass customization has emerged with features such as flexible machines, short production runs, frequent new product introductions, niche markets, highly skilled team workers, cross functional teams, low inventories, no defects, make to order, continuous learning, extensive communication with customers, empowerment and long-term trust-based relationship with suppliers. We thus see that, in order to survive in the new world, closely coordinated changes are to be made in a whole range of the firm's activities. Quick fixes such as buying a CNC machine centre, setting up e-mail facilities or buying CAD software tools or installing MRP–II will not help.

3. The manufacturing system

In this section, we present an integrated view of a manufacturing enterprise. Traditionally, manufacturing meant just the factory floor and during the 70's and 80's much effort spent around the world has aimed at improving technologies (FMS, CIM), practices (JIT, TQM), and effectiveness.

In this paper, a manufacturing system means the entire enterprise and consists of several interconnected subsystems, all of which have to act in cohesion so that customer-desired products are delivered on time. The subsystems include (see figure 1).

(i) *Suppliers*: They provide the subassemblies, components or raw materials just-in-time to the factory floor and play a critical role. Their flexibility, agility, defect control and organization structure should all be compatible with the goals, objectives and vision of the manufacturing system. For example, suppliers of companies introducing new products frequently, should also have similar capabilities. Having only a few loyal suppliers and effectively communicating to them the product and process designs are some of the new successful practices that Japanese firms have taught the rest of the world.

(ii) *Product development*: Being fast enough to be the first in the market with the right product, is worth more to the prosperity of most businesses than any other single manufacturing function. Aggressive global competition, emergence of new technologies, opportunities for high value quality of life products, and legislative and environmental requirements, are some of the drivers for new product development. Concurrent engineering, integrated product-process design, multi-functional teams,

and incorporation of customer voice, are some of the enabling solutions for effective product development

(iii) *Factory floor*: A flexible low-inertia factory, responding quickly to design and demand changes of the existing products and also to the need to produce new products, is mandatory. The layout, machines, work force and instrumentation and control hardware and software, are important elements of the factory floor. Recent trends include flexible manufacturing systems under computer control using client server architecture and agent-based scheduling strategies. Production planning and control strategies such as MRP-II also reside in the factory host computer.

(iv) *Integrated logistics*: Rapid response logistics to store and move raw materials, components, work-in-progress (WIP) and finished goods throughout the manufacturing system is another prime component of the manufacturing system. To maintain just-in-time deliveries, on time, to globally distributed customers and to make-to-order final products from subassemblies produced in several geographic locations, an agile logistic system that is mix- and volume-flexible, is essential.

(v) *Distributors*: The ultimate customer is served through the distributors and they play a crucial role in selling the product to the customer. They can also provide feedback on customer voice and their expectations. Direct marketing, telesales, contracts between distributors and manufacturers, outsourcing and strategic alliances are some of the issues here.

(vi) *Customers*: They are the ultimate users of products and services and receive the final output and generate revenue. Incorporating customer voice into the designs, delighting the customers with no irritants in the purchase and use of the products, gearing the organization towards delivering value to the customers, analyzing customers complaints and defections and fixing the process, are mandatory exercises.

(vii) *Organization structure*: Design of a highly responsive organization structure through which information and decisions flow through very quickly, manned by people who make decisions and effectively communicate among themselves and with other stake holders, is critical for success. The tendency nowadays is to replace the traditional bureaucratic organizational structure with team-based flat structures. Management of end-to-end customer business processes such as order-to-delivery, new product and process development, using multi-functional teams headed by process owner and aided by software agents is also a popular paradigm.

(viii) *Competition*: Staying close to competition, benchmarking their best practices, winning their core-customers are strategies for survival. Competitive intelligence is currently an art and involves inferring from publicly available data, the methods and practices that make a company number one.

(ix) *Technology*: This is the factor that pervades all the above subsystems. Successful manufacturing companies use advances in technology to introduce new products, to run the factories more efficiently and also to increase their customer reach. Electronics technology has had tremendous impact on the control, instrumentation and
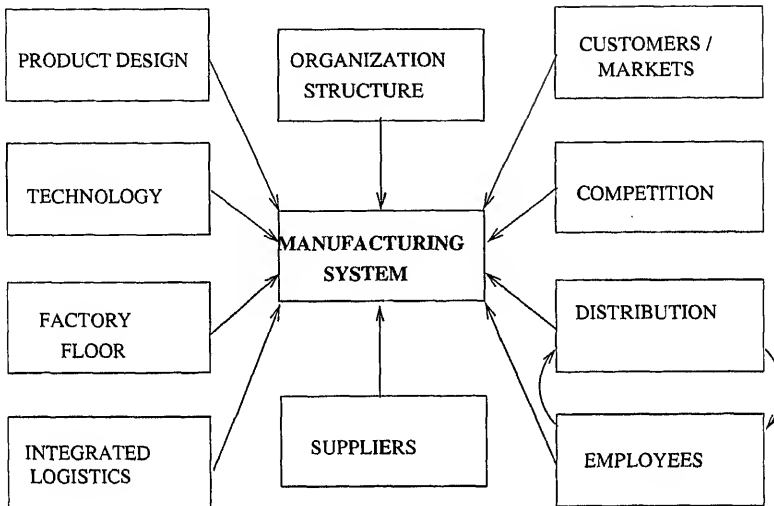
Figure 1. Competitive manufacturing as coordination of diverse skills and integration of multiple activities.

the way businesses are conducted (Venkataraman 1994). Internet, EDI, electronic funds transfer and e-mail have made several traditional systems obsolete. CAD systems have speeded up the design and retrieval process, and CIM has enabled flexible manufacture of high quality products. Managements should appreciate the value of technology and integrate it into their company's activities through effective use of R & D (Upton & McAfee 1996).

4. The business process

Traditionally manufacturing systems are viewed as a sequential arrangement of functions such as design, manufacture, R&D, marketing, finance, and so on. The recent trend is to view a manufacturing system not in terms of functions, divisions or products, but as a collection of value-delivering processes. Functional or hierarchical structures typically present responsibilities and reporting relationships whereas process structure is a dynamic view of how the organization delivers value to the customer.

Hierarchical organization structures based on functional divisions have several problems. Each function acts as a silo and hands over its output over the wall to the next function. Turfwars, dominance by functions such as finance and marketing, result in slow progress of work through the system. Lack of proper communication between functions results in work going back and forth with long iteration periods (see figure 2). The hierarchical arrangements in the functions require that decisions are to be sought from the top and work processes inch up and down the hierarchy ladder. Thus one finds that the ratio
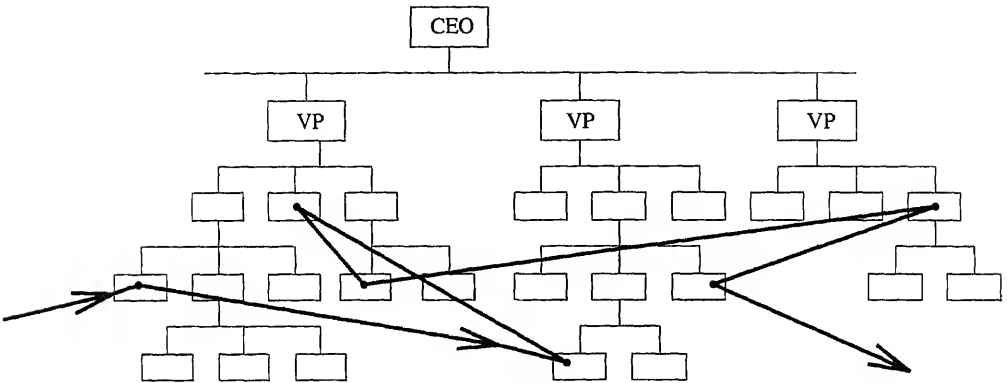
Figure 2.   Work flow through a functional organization.

cross-functional teams in a coordinated manner (Davenport 1993). Process thinking orig-
inated with the quality movement wherein emphasis is on process control and team work
rather than product inspection and command and control hierarchy. A process approach
to business implies heavy emphasis on *how* work is done rather than focus on specific
products (see figure 3).

DEFINITION 1
A process is a structured, measured set of activities ordered in time and space, designed
to produce a customer-desired output.

Clearly, structured processes are amenable to measurement in a variety of dimensions:
cycle time, defects, variability, flexibility etc. Examples of some typical processes include:
order-delivery process, new product development process, supply chain process, factory
floor process, maintenance process etc. A process perspective is a horizontal view of the
manufacturing system that cuts across the organization with product inputs at the beginning
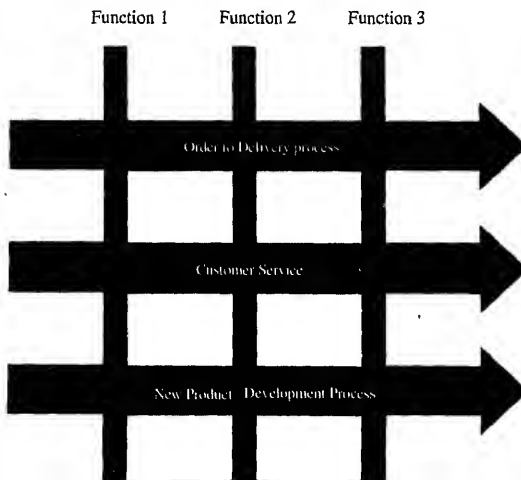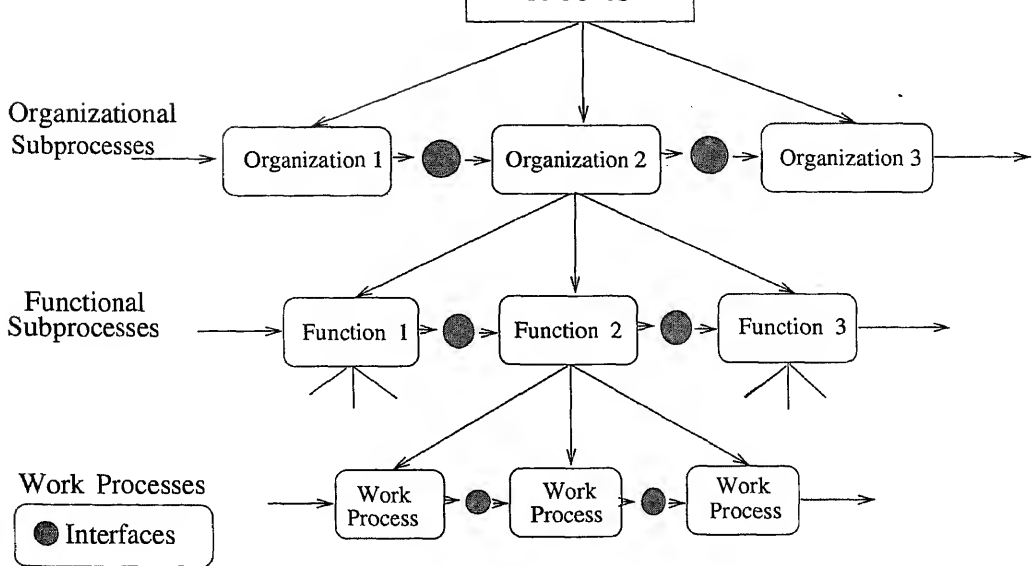
Figure 4. Business process hierarchy.

and customers at the end. Subscription to this view means deemphasizing the functional view of the business. The primary issue in vertical organization is the ill-management of hand-off between functions. Process orientation either eliminates hand-offs or coordinates them effectively. Processes are typically cross-functional.

There are several ways of decomposing a business process. We find it useful to decompose it into various subprocesses and work processes as shown in figure 4. There are four different types of sub-processes in a business process. They are the following.

(1) *Work process*: This is a value adding activity with well-defined inputs and outputs with someone responsible for it. Examples include machining by a machine centre, transportation from one location to another by a truck, bill processing by a clerk etc.

(2) *Interface between work processes*: Interfaces are white spaces between work processes and interface management involves documented procedures for transfer of work from one work process to another. Some of the procedures may involve automated transfer of the workpiece from one machine to another. Generally, work processes may be within the same function and hence management may not be difficult.

(3) *Functional interfaces*: These are the procedures to be followed when the output of a work process in a given function in an organization is transferred to another in a different function in the same organization. Interfaces between design, manufacturing, marketing etc., come under this category. These are smoothened out by using cross-functional teams to manage the subprocess in the organization and also by putting a measurement system in place that gives more weightage to the contributions made to the process by individuals and functions. One of the biggest achievements of process

orientation is recognizing the problems associated with functional interfaces and evolving procedures for managing them.

(4) *Organizational interfaces*: These are again relationships, procedures and activities performed when a process transits from one organization to another. Examples include supplier–manufacturer, manufacturer–distributor–retailer, distributor–customer relationships. Here again through certification and alliances one can smoothen the interfaces and avoid all non-value adding and time-consuming activities such as selecting the suppliers through quotation, redundant activities such as incoming inspection by customers and outgoing inspection by suppliers, financial guarantees, multi-sourcing for reliability reasons etc. Through partnerships one can avoid overhead costs as well as smoothen the workflow.

Generally, functional and organizational interfaces are unmanaged or managed infrequently by higher levels of management. The management procedures, when they exist, are outdated. The prescription is to treat the interfaces also as work processes, with an ownership, well-defined inputs and outputs, and monitor their performance through appropriate measures. The goal is to smoothen the workflow and make it as continuous as possible.

5.   Supply-chain and order-to-delivery processes

Supply chain process and the order-to-delivery processes are the most important business processes of a company and directly involve the customer. ODP starts with the arrival of an order from the customer and ends with the use of the product by the customer. In fact, some companies consider after sales support, and the return and recycling of the used product also as a part of the order-to-delivery process.

5.1   *Supply-chain process*

DEFINITION 2
Supply-chain process (SCP) encompasses the full range of intra-company and inter-company activities beginning with raw material procurement by independent suppliers, through manufacturing and distribution, and concluding with successful delivery of the product to the retailer or at times to the customer (Cohen & Lee 1988).

Supply-chain management is the coordination of the supply-chain process, i.e. integration of the activities/subprocesses involved in procuring, producing, delivering and maintaining products/services to the customer who are located in geographically different places (see figure 5).

Long term issues in SCP involve location of production, and inventory facilities, choice of alliance partners such as the suppliers, distributors, and the logistics chain. The long term decisions also include make-to-order or make-to-stock policies, degree of vertical integration, capacity decisions of various plants, amount of flexibility in each of the subsystems etc. (Connors *et al* 1995).
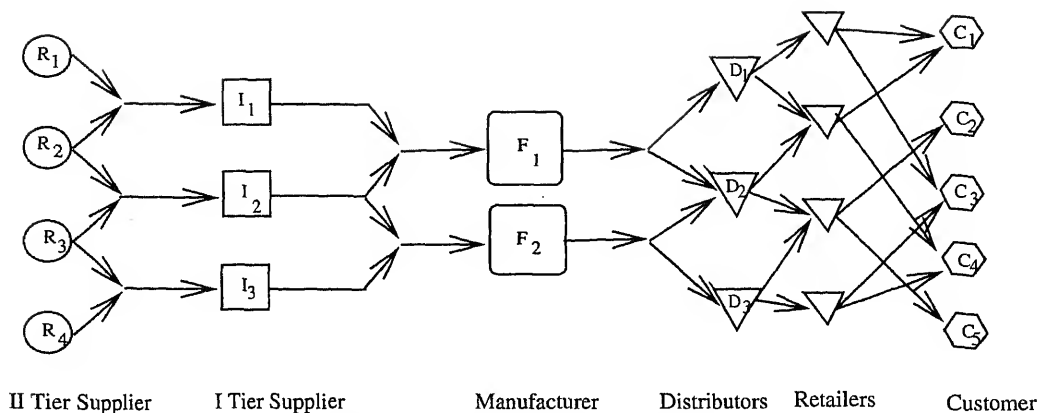
Figure 5. The supply-chain network.

| II Tier Supplier | I Tier Supplier | Manufacturer | Distributors | Retailers | Customer |

The operational issues in SCP are the ones we are more concerned with here. They include cycle time, on-time delivery, cost effectiveness, flexibility and quality. Identification of customers and triggers for their loyalties (cost, after sales service, on-time delivery) are also to be assessed.

5.2 *Order-to-delivery process*

DEFINITION 3

The order to delivery process consists of six steps starting from accepting orders through order configuration, sourcing the order, managing the order, monitoring and finally billing and cashing.

We briefly describe each of these functions below (see figure 6).

(1) *Accept orders*: Customers choose many ordering methods and most companies accept orders through traditional means like fax, mail or through sales representatives or through EDI, telemarketing or direct marketing channels. Orders can also be placed on the internet, provided the necessary infrastructure is available. Customers can then track the order on the internet. The order information should be visible to all members of the supply chain. One important issue in this step is order selection and prioritization. All orders need not be accepted and not all customers are equal. It is known that 80% of the orders come from 20% of the customers for 20% of the products. These customers should be treated as "sweet spots" and given priority.

In this step, we have four tasks,

(a) Order receipt.
(b) Order selection.
(c) Credit check: Previous customers are checked for default; new customers are checked for credit rating. A rule-based expert system can suggest credit limits. Any exceptions that may arise are handled manually.
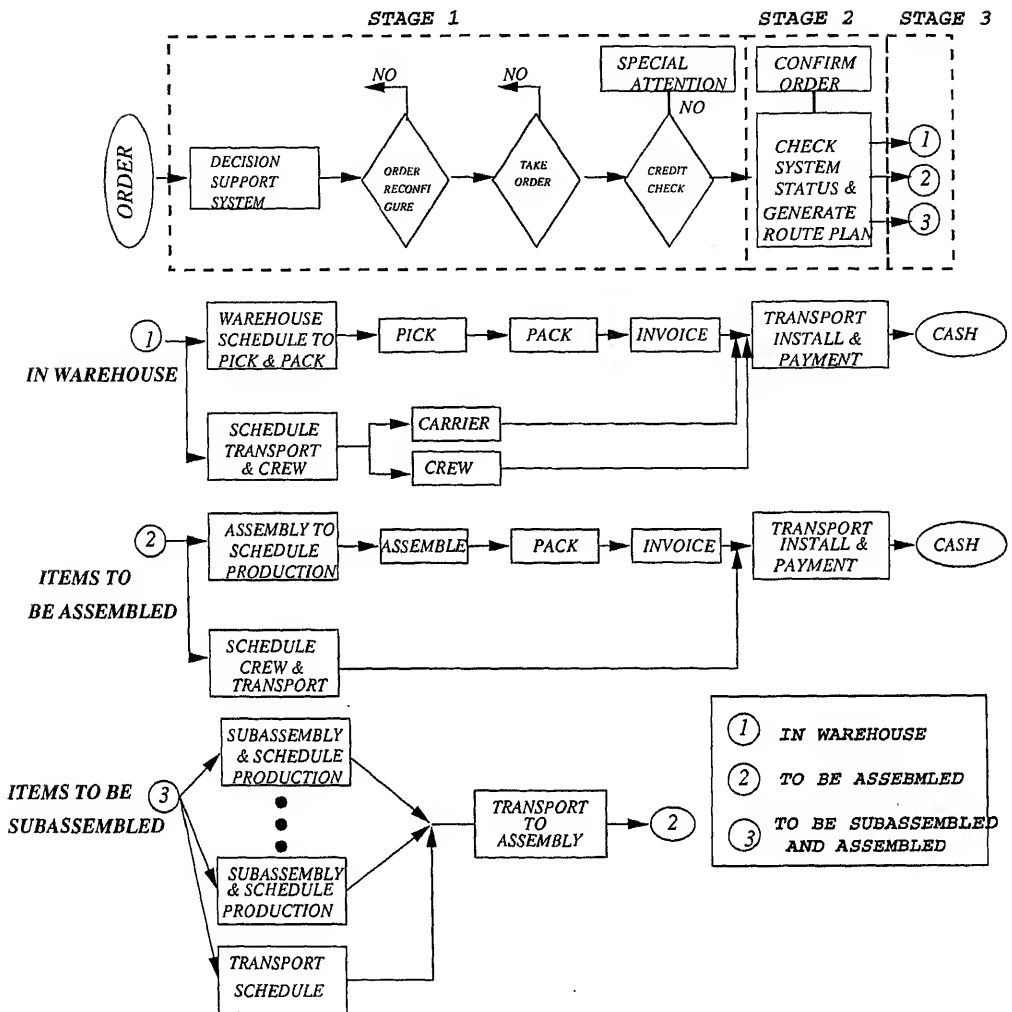
Figure 6. Order-to-delivery flow.

(d) Order confirmation: This involves the following.

- Check warehouse inventory and the allocation to current orders and back orders at all warehouse locations and determine if the current order can be met from nearby warehouses.
- Check the production orders in various plants.
- Make a decision on how the items from the order are to be shipped to the customer.
- Estimate cost and time.
- Dialogue with the customer on cost, delivery date and site planning if it involves equipment installations etc.
- Confirm the order, price and delivery date and time to the customer.

Figure 7 describes the controller interconnections that would enable the above functions to be performed.

IN WAREHOUSE ①			PICK/PACK	
ITEMS TO BE ASSEMBLED ②		ASSEMBLY	SHIP	
ITEMS TO BE SUBASSEMBLED ③	SUBASSEMBLY	ASSEMBLY		SHIP
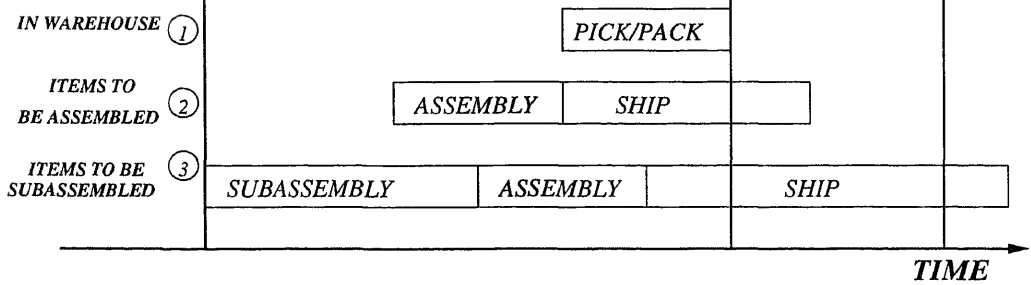
TIME

Figure 7. Delivery times for the three situations.

(2) *Configure order*: This consists of the following steps.

    (a) Identify the complete list of products and services that are contained in the order.

    (b) Plan and source each product and service in the order.

    (c) Synchronize delivery of each of the items with availability of service personnel and transport as well as the convenience of the customer.

(3) *Source the order*: If the order can be met from a nearby warehouse, then all activities relating to picking and packing are initiated. If the delivery time allows, orders may be sourced from a nearby manufacturing assembly plant by passing the local distribution. In this case, the order is included in the production schedule. If the plant does not have preassembled subassemblies, then they are sourced from assembly suppliers and is assembled and sent directly to the customer. The idea is to supply the order within the delivery window (see figure 7).

(4) *Order management*: This is the final step and involves the steps outlined below.

    (a) Making plans for coordinated delivery of the product to the customer.

    (b) Customer's order execution plan is communicated to each member in the supply chain responsible for shipping the product or providing a service such as transportation or installation. Planning for transportation can start before shipment is ready.

(5) *Order monitoring*: Like in project management, the progress of the order is monitored and actual progress is compared with the anticipated progress. The signals of completion or failure are sent to appropriate agencies which automatically monitor the progress.

(6) *Billing and cashing*: Invoices are certified after installation of equipment and payment authorized to customer's accounts payable, who transfers funds via EFT to company's accounts' receivable.

The above ODP can be automated using EDI, bar coding, decision support systems and other IT tools to achieve rapid cycle times and high customer satisfaction (figure 8).
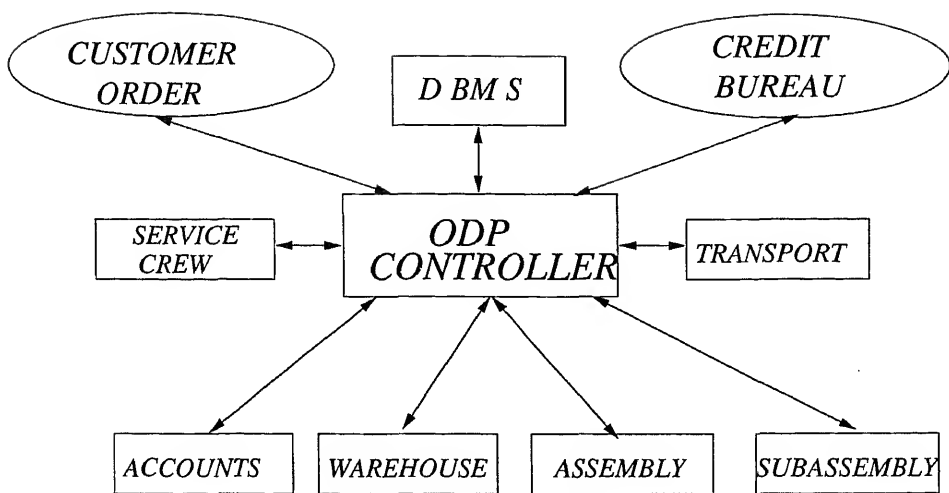
Figure 8. ODP controller.

6. Performance analysis: A process approach

Traditional performance analysis of manufacturing enterprises assumes that the enterprise is weakly coupled, i.e., the subsystems work almost independently. Paper-based communication, high inventories and arm-length relationships with suppliers, distributors and other stake holders justified this assumption. Performance analysis and optimization are conducted for individual subsystems and near-optimal performance of the total system was obtained by summing up the results. Each company is happy with its own operational and financial measures. Each function concentrates on and optimizes its own measures: sales on orders, manufacturing on throughput, distribution on shipments that maximize the truck utilization etc. This could result in sub-optimization at the overall systems and delayed customer deliveries. Products are designed for ease of manufacturing (DFM) and in assembly (DFA) and so on, but without worrying about logistic costs, import duty tariffs etc. In the name of reducing inventory costs, companies push the inventory either upstream to suppliers or downstream to distributors. Little do they realize that although they may have made cost reduction in the short term, the entire supply chain cost i.e., cost to the customer is the same or higher. Several intermediate manufacturers (e.g. automobile part suppliers) have to maintain lots of inventory because their customers (Original Equipment Manufacturers) have erratic ordering pattern, bad scheduling etc.; their suppliers are unreliable and are not quality conscious. Although they are lean and mean in their operations, they become heavy to keep the company efficient. Individual subsystem optimization is thus leading to total system sub-optimization.

Our thesis here is that a process-oriented approach to performance analysis and optimization is the correct one and all the ills mentioned above can be eliminated. Furthermore, demanding customers and global competition dictate that the entire chain, i.e., the supply chain or the order-to-delivery process should be effective and not just a link like manufacturing or marketing. In fact, future competition will be process based and it will be

the supply chain versus supply chain or new product development versus new product development, strategy versus strategy, etc.

Also, present day manufacturing systems are strongly coupled, both in terms of material and information flows. All stake holders in the manufacturing enterprises are connected via internet and intranet and also share a common database. Inventories are made very low and customers (both internal and external) insist on just-in-time deliveries. Further, suppliers, distributors and customers have a fundamentally different and cooperative relationship called partnership for sharing information, finances and helping one another in resolving quality and technology problems. Decisions in location of various facilities, manufacturing plants, distribution centres etc. are now taken by considering all costs: supplier-manufacturing, customer-manufacturing, distribution, transportation and other infrastructure costs. One may find that typically manufacturing cost accounts for only 20–30% of the total cost. These changes make the enterprise strongly coupled and any analysis has to be conducted for the entire monolithic system.

As discussed earlier, we consider a manufacturing system as a bundle of business processes. The business processes are value delivery streams starting with a customer need and ending with customer satisfaction. We now define the performance measures for business processes. For ease of presentation, we target all results from now on for the order to delivery and supply chain processes. The important performance measures of business processes are the following.

(1) *Lead time*: The lead time of a business process is the interval between start and end of the process. It is the concept to market time in the case of the product development process and the clock time between placing an order to the delivery at the customer site in the case of ODP and start of raw material ordering till the final assembly reaching the retailer in case of the supply-chain process. Lead-time reduction by removing non-value adding activities, and effectively managing interfaces with suppliers, manufacturing, logistics and distributors is an important exercise.

(2) *Customer service*: Delivery of customer-desired products at the right time, in the right place and in the right quantities every time is the goal of customer service. Customers could be external customers (or end users) or internal customers.

(3) *Dependability*: This is the reliability of product delivery and is an operational issue. It measures the ability to manage disruptions, such as machine failures, worker absenteeism, truck failure or supplier failures, rush orders etc.

(4) *Cost*: Like the lead time, cost also provides rich insights into process problems and inefficiencies. Interface costs, margins, costs in negotiations, inspection etc. are a waste and provide avenues for cost-cutting strategies.

(5) *Quality*: Quality is management of all the work processes so that they are on target with low variation. This is achieved through root-cause analysis of defects and failures and elimination of the causes of failures.

(6) *Flexibility*: Ability to meet customer requirements under various environmental uncertainties in various dimensions such as delivery time, schedules, design and demand changes etc. Flexibility of business processes is closely related to product structure or

(7) *Capacity*: This is the total output rate of the business process. All the work processes are to be balanced in capacity otherwise there will be bottlenecks and delays. Strategic alliances are common among various work process owners since vertical integration is almost impossible. A little amount of over-capacity to meet rush demands would improve operational measures.

(8) *Asset utilization*: The assets in some business processes (like the supply chain process and the order-to-delivery process) such as manufacturing plants, warehouses, communications infrastructure, fleet of vehicles etc. are worth billions of dollars at times. Their utilization is an important issue.

One could define other measures such as the throughput of the business process and the work-in-progress along the process. Also, product variety supplied and defects in the process are important measures which could be inferred from the above. We deal only with flexibility in this paper.

7. Flexibility

7.1 *Anatomy of changes*

Literature defines flexibility as a system's capability to cope effectively with a wide range of environmental changes and internal variations without deterioration in system performance in terms of cost, quality, lead time and on-time delivery. It is certainly a desirable virtue to possess for a manufacturing system in these times of global competition, turbulent changes and mass customization. It is certainly not a new concept and has been studied in the economic and organizational context. Earlier studies in flexible manufacturing were emphasizing flexibility in the context of the factory floor (Upton & McAfee 1994, 1995). In this section, in contrast, we are concerned with the flexibility of the entire business process.

Since a business process is an ordered set of work processes, it is essential that the work processes be flexible for the business process itself to be flexible. Flexibility enhances performance measures such as lead time, quality, and on-time delivery. It also allows the manufacturing system to cope with uncertainties such as those listed below.

(i) *Resource changes*: Variations in the number of human and machine resources on a factory floor due to machine failures, absenteeism etc; transport breakdown in the logistic system; a rush order from a valued customer; supplier bankruptcy etc. are issues that arise routinely. Process management should be able to cope with these kinds of change.

(ii) *Design and demand changes in the product*: The changes could be either planned or unplanned. Customer demand is random and this combined with inaccurate forecasting will cause uncertainty in the design and mix of products. Also, proactive introduction of new products to beat the competition will reflect as planned change.

(iii) *Technology changes:* These could be continuous or discontinuous. Discontinuous technology changes such as the ones in the PC industry and the hard disk drives are

difficult to cope with. The company should have the ability to predict and develop competences in new technologies for future product generations and also have the capability to evaluate the risk associated with new ventures.

(iv) *Socio-political changes*: Deregulation of telecommunications, airlines, information networks etc. has a big impact on existing players. Liberalization of certain closed economies had the same impact. Legislations on health care had tremendous impact on hospital and health insurance systems.

7.2 Coping with uncertainty

We see that as we traverse from resource changes to design and demand changes in the product to socio-political changes, both the magnitude and the effect of change will increase. They can be classified as operational, tactical, and strategic changes and the flexibility strategies could be correspondingly named. The resource changes occur daily and appropriate procedures have to be designed and built into the system. Events such as failure of a machine or a truck or a rush order occur at random times and places. Procedures, like those in hospitals and civil defense, should be evolved for all frequently occurring events. Design and demand changes occur, say monthly, and they have to be met through proper scheduling of orders. They are tactical in nature and involve suppliers, logistics and distributors. The technology changes are sporadic but occur in a predictable way for companies with learning capabilities. Some companies see the opportunity and lead the change in an offensive way. They are strategic in nature and involve proactive strategies for product development, technology adaptation, etc. Socio-political changes are outside their control but some companies turn these into opportunities as well.

The flexibilities built into the enterprise should have the abilities to cope with these changes. These abilities are built into the enterprise via technology, procedures, control mechanisms such as scheduling, information processing etc. It is important to analyze the changes that a system is subjected to, so that appropriate flexibility strategies could be designed and implemented.

The best way to cope with uncertainty is to eliminate the sources of uncertainty. This may not always be possible but one can certainly reduce the amount of change. For example, customer surveys and monitoring point-of-sale information, redesigning product range to increase part-commonality and delaying final assembly until receipt of order, will reduce demand uncertainty; preventive maintenance, use of diagnostic expert systems, and built-in fault tolerance will reduce the down-time and increase the system availability; quick adoption of new technologies such as IT and their proper implementation will remove all non-value adding activities and cut costs and delivery time to a minimum. Then one can confidently face, in fact, lead price wars in the face of deregulation.

At times, it would help eliminate the need for change by replacing a rough terrain with a smooth road rather than trying to cope with it. By providing a smooth road, and a constant environment for a vehicle, one avoided the need for flexible legs. By providing transfer mechanism in an automobile assembly, which can transfer the workpiece across machine tools, one has eliminated the need for sensory and manipulative human functions (Viswanadham & Narahari 1992).

Group technology cells, each focusing on a product family, streamlined the product flow on the factory floor thus eliminating the need for sophisticated material tracking and scheduling algorithms.

A flexible manufacturing enterprise whose business processes can cope with all the above changes will be extremely complex, expensive and time-consuming to install. It requires redundancy in terms of excess capacity, space and time which again will increase the cost of the products. In an environment of fast-changing technologies, shrinking product life cycles, and changing customer attitudes, one should strike a balance between cost and time for implementation and hedge against near-term uncertainties. One has to recognize that each industry/firm has only a finite life and eventually has to open new factories for new products and phase out old ones.

An analysis of the product line of the enterprise for the next five years using good forecasting techniques, is generally done to determine the uncertainty in demands and the appropriate manufacturing system (focused cellular manufacturing cells, flexible manufacturing system or flexible transfer line) is decided. The kind of flexibilities that are needed to meet the customer demand could also be determined at this stage. Methods such as strategic alliances, acquisitions, out-sourcing, subcontracting are used to augment the competences. Benchmarking is used to determine the best in class manufacturing practices and the performance gaps.

Thus we see that a preliminary analysis of the uncertainties will lead to an appropriate manufacturing system configuration which can effectively cope with their changes induced by them.

7.3   Flexibility and technology

Information and communication technologies such as EDI, EFT, CAD, FMS, ASRS, groupware, local area networking, bar coding, internet, intranet are extensively used to speed up the process as well as to deliver better value to the customer. Automation technologies such as FMS, ASRS, MHS are very expensive, others such as EDI and bar coding require process simplification and analysis before implementation. The communication technologies such as EDI, EFT etc. are integrative technologies and their added power comes from the fact that information collection, storage and processing could be done in real-time and transferred to all the stake holders. For example, order information, point of sale and inventory status could be monitored by all concerned such as distributors, manufacturers, suppliers etc. This information can be gainfully employed to schedule production based on the needs of the customer. Advantages and use of all the technologies mentioned above is well documented in the literature (Corline & Essaides 1993; Hartley 1993).

All the core business processes such as the supply chain process, order-to-delivery process, new product development process, factory floor process and logistics process are all influenced by the above technologies to improve the cycle time, quality and enhance the delivery reliability (Hartley 1993). Quick changeovers, sharing of information (on customer order, product designs, inventory status, point of sale), supplier, logistics and distributor partnerships are enablers of high performance, low inventories and high flexibility (Corline & Essaides 1993; Srinivasan *et al* 1994). Customer desired product variety can be supplied without penalty on performance (Kekre & Srinivasan 1990). Low inventories result

buy rigid or flexible technologies for a given demand forecast, economics and advantages of EDI, whether a supplier should adopt EDI, whether a customer should invest in a supplier's infrastructure (production technologies, EDI, etc) (Wang & Seidmann 1995), vertical integration or virtual integration, whether a supplier should follow a leader model or follower model etc. (Srinivasan *et al* 1994; Wang & Seidmann 1995).

Technology has made possible the sharing of various kinds of information and partnering. Virtual integration is an alternative to vertical integration and it provides flexibility to meet changing demands. In a changing scenario, whether to change the partnership or invest in enhancing the capabilities of existing suppliers is the big question. Much can be said on both sides. Also, choice of suppliers, distributors and transporters is an irreversible decision until the end of the contract. Changing the partners involves considerable time and effort and is comparable to the set-up time and cost. It is important to note that good partnerships will cut both cost and cycle time. A manufacturer can maintain the shelf space in a supermarket bearing the inventory, transportation and obsolescence costs in return for point-of-sale information (Venkataraman 1994). A distributor can maintain the inventory of hot spares and medical supplies to have big customers locked-in. Bennetton, Walmart, and Dupont have used technology to redesign their business processes for high flexibility in terms of response to customers and high performance in terms of low cycle times (Corline & Essaides 1993; Hartley 1993).

7.4 *Flexibility in manufaturing enterprises*

It is common misunderstanding that flexibility is achieved through flexible machine or computer hardware acquisition. The truth is far from that. Indeed several companies incurred losses, let alone breaking even, because productivity reduced with introduction of new manufacturing hardware. Process variety complicates the parts supply and assembly process because more parts require a greater coordination to get the right part into the worker's hands at the exact instant the guided vehicle brings the car to the worker's station. Because of the complexity induced by variety, many companies view flexibility management as a necessary evil (Sethi & Sethi 1990).

Manufacturing enterprises increasingly look like restaurant chains. Customers place their orders, waiters transmit the specification to the kitchen and a team of cooks assemble the product. The products are designed already and part programs are available, as soon as customers order from a table look-up, products are scheduled and delivered. This kind of flexibility is *static flexibility* or product-centric view of flexibility. On the other hand, *dynamic flexibility* is creation of capability to act in response to opportunities as they arise over time. Competences to developing new designs and new products to manufaturing customer desired products and to deliver them faster than competition are tenets of dynamic flexibility (Macduffie *et al* 1996).

There are four basic types of flexibility: mix, volume, new-product and delivery time flexibilities. See table 1. Each is important in a different environment (Suarez *et al* 1996). These flexibilities are implemented through a variety of factors such as production technology,

Table 1. Types of flexibility.

Mix Flexibility	Ability of a system to simultaneously produce a number of different products in a given period.
Volume Flexibility	Ability of a system to change significantly the production level and the composition of the product mix in a short time span.
New Product Flexibility	Ability of a system to add or substitute new products to the product mix over time.
Delivery Time Flexibility	Ability of a system to reduce the order-to-delivery time

product management techniques, relationship with suppliers and distributors, human resource management and product design. It is important to realize that different types of flexibilities are important in different competitive situations. For example, mix flexibility is important when a firm has a broad product line and caters to different market segments. There are several ways of achieving each type of flexibility. Mix flexibility may be achieved through skilled workers or programmable equipment. New product flexibility is needed in technology intensive markets. Volume flexibility is important in volatile markets.

Flexibility management is a competence that involves skillfully managing several resources of the system including automation hardware, software, people, organization structure, suppliers, customers, distribution channels, and factory floor control systems. It is an integrity-related competence and involves collective learning in the organization, coordinating diverse production skills and integrating multiple streams of technologies. It is a capability for deploying various resources of the company using the organizational processes to efficiently and economically produce a wide variety of part types. We elaborate this point with respect to the relationship with suppliers and distributors.

A formal relationship between suppliers and distributors is essential for a positively correlated flexibility productivity relationship. First, the capability of a manufacturer to offer a rich variety of products is dependent on the supplier's capability to produce a variety of component parts, i.e. the supplier's flexibility in several dimensions: delivery time, mix, volume and new products. If components for each product in the mix are sourced from different suppliers, then the management overhead increases enormously. Secondly, when a plant has machine failure problems or when there are sudden volume surges time-sensitive orders can be sub-contracted to dependable contractors.

The ability of the manufacturing system to produce a variety of products should be matched with the ability of the distributor to pass the variety on to the customer by proper advertising and maintaining appropriate inventory levels. Also, distributors by their close interaction with customers can easily identify their true needs and preferences so that the company can *produce what sells* rather than trying to *sell what is produced*, thus minimizing "market defects", i.e. producing things that customers do not want. One cannot underestimate the influence of distributors in variety management, both in information collection and also in sales. Further, the use of information technology tools such as electronic data interchange, electronic funds transfer and customer sales tracking systems by dealers and suppliers enhance the delivery time flexibility.

From the above discussion it is clear that flexibility needs to be defined, designed and created for end-to-end business processes.

8. Flexibility in business processes

DEFINITION 4
A business process is flexible if it can effectively manage or react to change with little penalty in time, cost, quality, or performance.

This is of course a very general and abstract definition. Our discussion in this section will concentrate on issues concerning any business process; the specific concerns of new product development, supply chain and order-to-delivery process will be dealt with as examples. As we saw earlier, a business process is an ordered set of work processes and interfaces (both functional and organizational). The interfaces are generally managerial processes such as sharing information, rules and procedures etc., depending on relationships with suppliers, distributors, customers etc. Work processes typically involve tasks such as design, manufacturing, marketing, sales etc. It is essential that the interfaces are smoothened out and work processes are flexible for the entire business process to be flexible. Interfaces between functions and organizations are like setups which have to be reduced by continuous improvement.

Previous studies on flexibility have so far been confined to factory floor technology, setup times, WIP, and so on. The volume and mix flexibilities are defined with the factory floor in mind. Here we talk of the entire business process, as flexibility of some of these processes such as customer acquisition may not include the factory floor. Flexibility in human resources, infrastructure flexibility and flexible management structures are all examples of flexible processes beyond the factory floor.

Now we consider SCP and ODP and discuss flexibility issues in these processes.

8.1 *Supply chain process*

DEFINITION 5
A flexible supply chain process is one that responds effectively to changes in volume, product mix, delivery times and delivery routes without deterioration in cost, quality and lead time.

It is essential all subsystems be flexible for the supply chain process to be flexible. Flexibility management is a capability that has to be built up over time through the use of skilled work force, automated equipment, IT tools, computer control systems, bench marking and implementing the best practice and the like.

We will first discuss the various types of flexibilities for the supply chain. Essentially flexible supply chains accommodate special customer requirements, provide customized service, allow product modification while the order is in process, introduce new design features and so on.

8.1a *Volume flexibility in supply chain process*:

DEFINITION 5.1
A supply chain is *volume* flexible if a customer order with different product mixes and volume levels can be processed for rapid delivery.

It is essential that small batches of products are produced and delivered for a system to be volume-flexible. This implies that setup times are small all along the process. This is because the economic batch size in any work process (manufacturing or transport) is an integral multiple of its successor.

8.1b *Mix flexibility in supply chain process*:

DEFINITION 5.2
A supply chain is mix flexible if the system can produce a number of products simultaneously and deliver them to the customers.

This capability indicates the breadth of the product line and ability for quick change-overs. The suppliers are either mix-flexible or there are more number of suppliers. Also warehouse and transportation should be able to handle different sizes, shapes and installation procedures (multi-skilled labour).

Excessive product variety induces several problems in both performance and management. The system complexity increases with a greater number of suppliers (at least two or more for each component), since establishing partnerships, sharing information, helping in quality control, reducing changeover times etc. are all time- and effort-consuming processes. Thus variety means more design, more production planning and control, and more forecasting and more leftovers. While no one can disagree that one should have variety, it is necessary to find and manufacture the twenty percent of the products that win in the market.

8.1c *Routing flexibility in supply chain process*:

DEFINITION 5.3
This is the ability of the supply chain to produce and deliver to the customer through alternate routes or, equivalently, each function (manufacturer, warehousing, transporting) could be performed in more than one location.

Routes to supply equipment or to fill in the order can be ordinarily fixed but can be changed in the event of problems such as breakdown. Routing flexibility is generally obtained by duplicating each function in various locations, having over-capacity and redundancy in transportation, and efficient scheduling and control software. The average number of possible ways in which an order can be filled could be used as a possible routing flexibility measure. For example, an order for a workstation from an Indian customer, can be filled either from Singapore or Europe or USA in a variety of routes. Depending on the time available, it is sent by air freight or by ship.

8.1d *Delivery time flexibility in supply chain process*:

DEFINITION 5.4
A supply chain process is *delivery time* flexible if it can reduce or expand the delivery time as per customer requirements.

Here again, rush orders and delayed shipment requests are common from customers. Ability to reschedule the orders all along the supply chain, low variation of the lead times of all the work processes, quick change-over times, excess capacity in all resources, are some of the requirements for delivery time flexibility.

8.1e *New product flexibility in supply chain process*:

DEFINITION 5.5
A supply chain process is *new product* flexible if generations of several new products can be rapidly designed and marketed simultaneously.

This is a very important flexibility for most manufacturing companies. A cross-functional team with a process owner manages the product development process with reviews and testing at intermediate points in the process to maintain design quality. The team has members from among suppliers, distributors and at times even customers, apart from having functions within the organization, so that the designed products are manufacturable, saleable, and satisfy customer specifications. Time-to-market is critical for this process to gain the advantages of first mover (Ulrich & Eppinger 1995). Patenting, navigating through regulatory agencies, production, marketing etc. augment product development capability. For example, pharmaceutical companies are most R & D intensive and the high cost of drug production mandates introduction of the drug worldwide to make the most from the effort. Thus a company should not only have capability to innovate new drugs but also have downstream capabilities to navigate it through regulatory mechanisms, and for manufacturing, marketing, and distribution in domestic and foreign markets.

8.2 *Flexibility in order-to-delivery process*

We now consider the flexibilities connected with order-to-delivery process.

DEFINITION 6
An order-to-delivery process (ODP) is volume-flexible if a number of customer orders with different product mix and volume levels can be simultaneously processed for rapid delivery.

Mix, routing and delivery time flexibilities of an ODP can be similarly defined. Notice that volume flexibility of the ODP requires flexible order processing, and a flexible supply chain process with flexible manufacturing, logistics chain, marketing channels, etc. Also all these subsystems need not be within the boundary of a single firm, they could be part of a supplier-manufacturer-distributor value chain. For example, a catalog store has an ODP that is mix- and volume-flexible if it can reliably coordinate the supply chain. Similarly a car dealer is mix-flexible if he can arrange for delivery of any customer-desired car as quickly as possible. An ODP is routing-flexible if it has redundant suppliers, manufacturers, distribution etc. in various locations. Table 2 gives some measures of the flexibility for an ODP.

Table 2.   Flexibility measures for ODP.

Type	Measures
Mix flexibility	• Number of different products that can be supplied. • Optimistic changeover times and costs among different products (function of scheduling)
Volume flexibility	• Stability of cost of delivery over varying levels of production volumes. • Smallest profitable volumes of operation.
Routing flexibility	• Average number of ways in which a product can be ordered, manufactured, and delivered. • Average delay due to subsystem failures.

Three aspects of ODP flexibility – Flexibility is the response of a firm or a group of firms to uncertainty. We have identified three different types of changes: short-term, mid-term and long-term. Business processes and all systems involved in it have to manage these changes effectively. The built-in hardware, software and management procedures should be able to counter the variations or changes. We identify operational, tactical and strategic flexibilities in the ODP context.

8.2a   *Operational flexibility*:   Short-term changes occur very frequently, may be every few hours to every day. We assume that the organizational and functional interfaces are all smoothened out, i.e. coordination of deliveries with suppliers, collaborative arrangement for distribution, are all in place. A business process is operationally flexible if it permits a high degree of monitoring and control to accommodate short-term changes.

Basically operationally flexible firms have the list of possible contingencies that can occur or have previously occurred and a list of methods to deal with the situations. Like the emergence of procedures in civil, defence or medical fields, the approach taken can be reactive or for eliminating the root cause of the contingency. In either case the customer's problems are resolved. A firm should have the formal procedure in place to automatically accommodate routine or anticipated events occurring at random points in time, with staff empowered to make decisions. Encouraging creative response by staff will help in operational flexibility of processes such as ODP which are characterized by massive detail over a large geographical area. Use of outside service providers and strategic partners is also common. Some examples include:

(1) withdrawal of a the drug from the market because of possible contamination in a batch;

(2) failure of earth-moving equipment in a remote location, for which spares need to be delivered as soon as possible.

8.2b   *Tactical flexibility*:   An important characteristic of tactically flexible processes is the ability to switch quickly and cheaply between products. Real time data collection at point of sale, networked organization structure, EDI interconnection, bar coding etc. are important ingredients of the system. Good implementations include quick response manufacturing by Dupont and desktop publishing facilities by McGraw-Hill. Basically, in case of tactical flexibility, we are concerned with delivering known products to customers

subject to the required delivery schedule (both in quantity and time). Reduced setup times and costs at various machines in the suppliers' and host's manufacturing plants, fast communications, say through EDI, among all stake-holders, reducing transport costs either by using small trucks or by sharing the truck among different suppliers, packaging products in a form directly usable by the customer, having surplus capacity in the manufacturing and transport systems to take care of breakdowns in machines and transport, fault-tolerant control system to manage the system, are some of the attributes of such systems. Benetton, TESCO, and Walmart are examples of successful tactically flexible order-to-delivery business processes. Similarly, custom-made book publishing by McGraw-Hill is also an example of tactical flexibility in various business processes. In all these cases the changes are in the demands of the customer. The response of the company or group of companies to provide the range of products has to be quick. This is achieved through strategic networking and communication of suppliers and customers in case of Benetton, TESCO and K-Mart.

8.2c *Strategic flexibility*: This involves development of competence in technologies to lead the industry in all work processes and innovation in managing/eliminating interfaces. It is concerned with the ability to make good use of undisclosed opportunities, either through the production process or the product. It is the ability to respond to unmeasurable changes in the market conditions and unprogrammed advances in technology.

Any type of flexibility arises from two sources: from assets (tangible or intangible objects useful in product delivery) and organization or coordination of these assets to gain competitive advantage. Strategic flexibility is the ability to deploy the assets and capabilities to create, produce and offer new products to markets in a changing environment, dynamic markets and shifting technologies.

9. Product structure and flexibility in supply chains

Performance and flexibility issues in supply chain and order-to-delivery processes are intimately related to the product structure and also to the logistics network organization. Modular designs with customization occurring as late as possible in the production are preferred. Also manufacturing facilities are staged with customization occurring with the help of local suppliers to meet local language, power, and communication standards. The combined product and logistic network design yields the following structures (Williams 1981; Macbeth & Ferguson 1994):

(i) A straight-line interconnection of plants, each adding modules to the semi-finished product sent by the previous plant and passing it on to the successor. The product variety is limited, the material flow unidirectional and the decisions are infrequently made as to when to stop the line for switchover to another product type or how large the batch size should be. Mass production and continuous manufacturing are examples of this type. Batch sizes are large in this case.
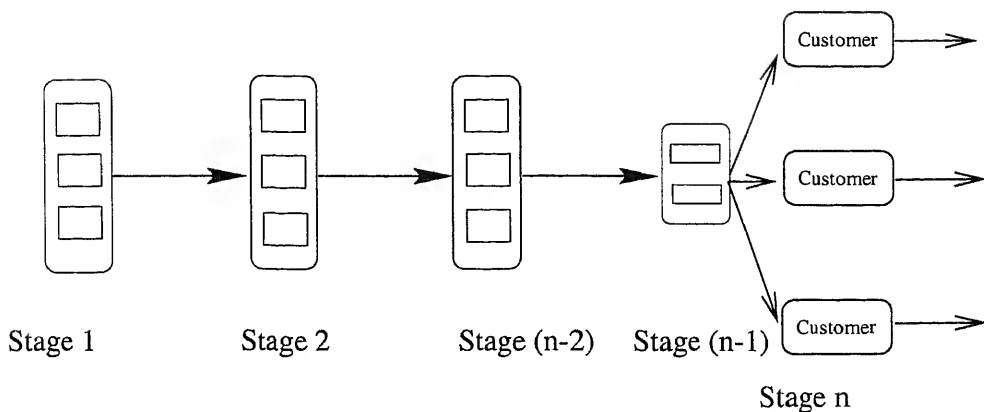
Figure 9.   Product (manufacturing plant) structure with late customization.

the customer end or at a local manufacturing plant or distribution centre. The structure of the plant looks similar to the one shown in figure 9.

Personal computers, IC chips, disk drives, laser printers and other electronic equipment follow this pattern. Here a product variety can be obtained through dedicated plants to perform the last stages of manufacturing. Also, modular design of the product keeping in view the logistic costs will reduce the total logistic costs as well as help in effective supply chain management. The individual customer demands could be highly variable, but the total semi-finished product demand variability is low. Hence it makes lot of sense to maintain inventories at the subassembly level which are customized based on individual demands and thus maintain low lead times (Lee & Feitzinger 1996).

The various stages of manufacturing could be dispersed across continents and transportation is often on the high seas. Thus the economic transportation batch size is larger than in closely located plants. This dictates that the production batch size of the predecessor plants also be high. Thus batch sizes and inventories are high and product variety is low in all the first $(n - 1)$ stages. The final stage, however, has dedicated low volume plants.

In the plant structure shown in figure 9, uncertainties in customer demands and long transportation times are basically met through staged manufacturing plants with dedicated technology, inventories, and large-to-medium batch sizes. Supplier management, partnership with logistics agents, modular product designs are some of the enablers for good flexibility management in this case.

(iii) We consider the third case of plants where customization starts early in the production stages and which have a diverging architecture (see figure 10). Starting with a limited number of raw materials, a wide variety of finished products are produced. Such examples can be found in electro-mechanical systems such as motors, textiles, metal fabrication, and chemicals.

The enterprise is very complex from a managerial viewpoint. Several plants with similar manufacturing capabilities have to be managed, and have to maintain

Stages: 1   2   3   4
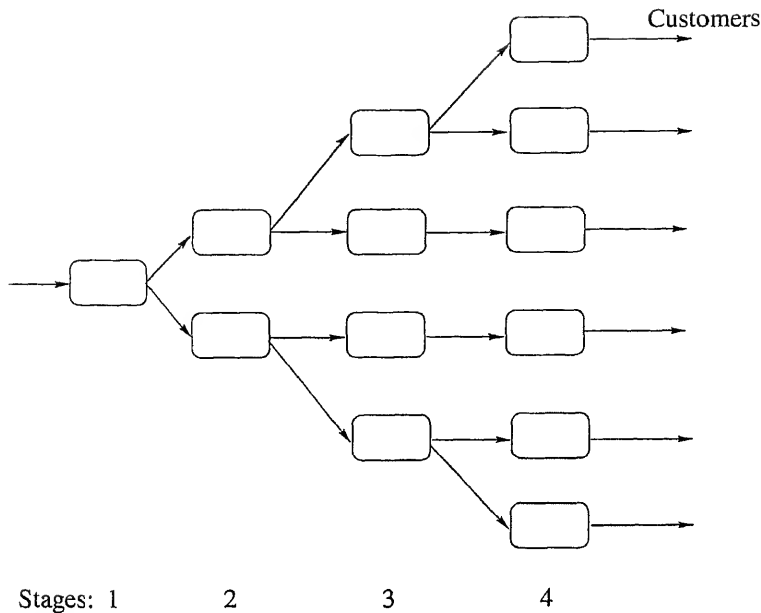
Figure 10. Product flow and manufacturing plant structure with early customization.

partnership relations with several suppliers, distributors, and customers. The plants could be staged or can be at central locations and made flexible using automated technologies. The batch sizes are typically low, and lead times are to be small. Low inventories are to be maintained for cost competitiveness. Integration using EDI and bar coding, and scheduling customer orders using point-of-sale information would increase the effectiveness of the supply chain. The mix flexibility for the supply chain is very high. Volume fluctuations are handled through large number of suppliers and quick feedback for point-of-sale information. The order has to travel through lots of organizations and the interfaces are generally not managed; instead high inventories are maintained.

(iv) The final architecture is assembly-oriented with converging architecture. One finds such patterns in the manufacture of aircraft and in construction. Numerous raw materials are transformed into sub-assemblies and finally into a huge assembly (see figure 11). Large number of mid-volume components are produced at various stages of manufacturing. Components are globally sourced from a variety of suppliers. The management is complex. Here again supplier management, flexible manufacturing, and information integration would enhance the enterprise's performance. Small batch sizes, sometimes one of a kind, and low inventories are typical here. The supply chain is definitely mix-flexible, for e.g., aircraft manufacturers like Boeing have wide-body, narrow-body aircraft and the 777's. The product mix gets smaller as the structure converges, high at the components level and low at the final assembly level. Scheduling to deliver products on-time is an important issue in both the construction and the aircraft industry.
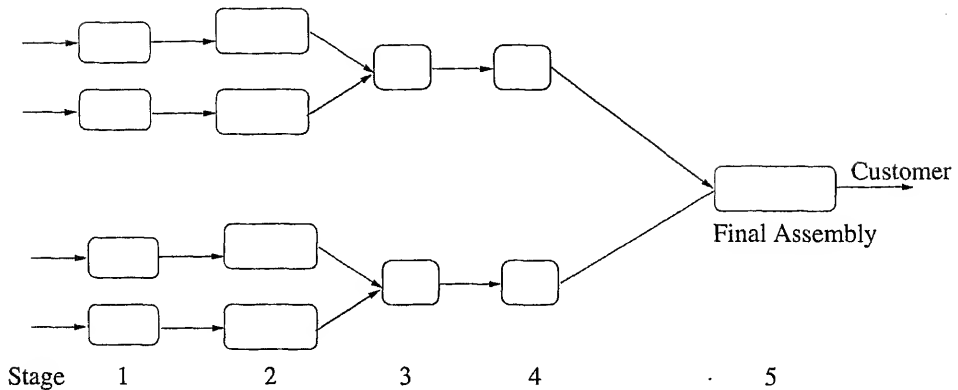
Figure 11.   Converging product (manufacturing plant) structure.

We have described a few typical architectures and used these as vehicles to discuss the flexibility issue. Our intention here is to point out the fact that flexibility, as the final user sees it in terms of variety, can be obtained in several ways depending on the product structure.

10.   Conclusions

In this paper, we attempted a complete definition of a manufacturing system as a set of integrated activities involving suppliers, production, distributors, customers, competition, logistics etc. and emphasized the need for coordination among these activities via management of end-to-end business processes. We also brought out the fact that a preliminary analysis of the uncertainties that a manufacturing system faces, and the subsequent design of the system for flexibility and also the design of control systems to reduce the effects of these uncertainties on system performance are essential for achieving sustainable competitive advantage. We defined the performance metrics for business processes.

We have defined flexibility of business processes and presented definitions and measures for important processes such as product development, supply-chain and order-to-delivery processes. Finally, the discussion presented here is equally valid for continuous and chemical manufacturing systems as well. Except for a manufacturing plant that has continuous dynamics, the rest of the system is identical to the one discussed here. Even in continuous processes, there is too much attention paid to the manufacturing process and less to logistics and the interface activities between various functions and organizations. Attractive gains can be made using IT and process management techniques in reducing the cycle time and improving quality in these processes.

References

Cohen M A, Lee H L 1988 Strategic analysis of integrated production distribution systems: models and methods. *Oper. Res.* 36: 216–228

Connors D, An C, Buckley S, Feigin G, Levas A, Nayak N, Petrakian R, Srinivasan R 1995 Dynamic modeling of re-engineering supply chains, IBM Research Report 19944

Corline D N, Essaides G 1993 *Time-based competition.* The Economic Intelligence Unit Research Report S-159

Davenport T H 1993 *Process innovation* (Cambridge, MA: Harvard Business School Press)

Hartley J 1993 *Electronic date interchange, gateway to world class supply chain management.* The Economic Intelligence Unit Research Report P-660

Kekre S, Srinivasan K 1990 Broader product line: a necessity to achieve success. *Manage. Sci.* 36: 1216–1231

Lee H L, Feitzinger E 1996 Product configuration and postponement for supply chain efficiency. Technical Report, Stanford University, Stanford, CA

Macbeth D K, Ferguson N 1994 *Partnership sourcing: An integrated supply chain management approach* (Pitman)

Macduffie J P, Sethuraman K, Fisher M L 1996 Product variety and manufacturing performance: evidence from international automotive assembly plant study. *Manage. Sci.* 42: 350–369

Milgrom P, Roberts J 1990 The economics of modern manufacturing: Technology, strategy and organization. *Am. Econ. Rev.* 80: 11–528

Prahlad C K, Hamel G 1990 The core competence of the corporation. *Harvard Business Rev.* May–June: 79–91

Sethi A K, Sethi S P 1990 Flexibility in manufacturing: A survey. *Int. J. Flexible Manuf. Syst.* 2: 289–328

Smith B 1993 Six-sigma design. *IEEE Spectrum* 30(9): 43–47

Srinivasan K, Kekre S, Mukhopadhyay T 1994 Impact of electronic data interchange technology on JIT shipments. *Manage. Sci.* 40: 000–000

Stalk G N Jr, Hout T M 1990 *Competing against time* (Free Press)

Suarez F F, Cusumano M A, Fine C H 1996 Flexibility and performance: A literature critique and strategic framework. *Manage. Sci.* 44: 223–240

Ulrich K T, Eppinger S D 1995 *Product design and development* (New York: McGraw-Hill)

Upton D M, McAfee A 1994 The management of flexibility. *California Manage. Rev.* Winter: 72–97

Upton D M, McAfee A 1995 What really makes factories flexible. *Harvard Business Rev.* July–August: 74–84

Upton D M, McAfee A 1996 The real virtual factory. *Harvard Business Rev.* July–August: 123–133

Venkataraman N 1994 IT established business transition: From automation to business scope redefinition. *Sloan Manage. Rev.* Winter: 000–000

Viswanadham N, Narahari Y 1992 *Performance modeling of automated manufacturing systems* (Englewood Cliffs, NJ: Prentice–Hall)

Viswanadham N 1996 Competitive Manufacturing Enterprises. *India 2010 – Leadership through science and technology* (Bangalore: Indian Academy of Sciences) (under preparation)

Wang E T G, Seidmann A 1995 Electronic date interchange: Competitive externalities and strategic implementation policies. *Manage. Sci.* 41: 000–000

Williams J F 1981 Heuristic techniques for simultaneous scheduling of production and distribution in multi-echelon structures: Theory and empirical comparisons. *Manage. Sci.* 27: 000–000

Womack J P, Jones D T, Ross D 1990 *The machine that changed the world* (Harper Perennial)

Manufacturing supply chain modelling and reengineering

KUMAR BHASKARAN and YING TAT LEUNG

IBM Corporation, Research Division, Thomas J. Watson Research Center,
P.O. Box 218, Yorktown Heights, NY 10598, USA
e-mail: [bha,ytl]@watson.ibm.com

Abstract.   The first-wave of reengineering, during the first half of the nineties, focused on making organizational changes and used primarily *information models of supply chains* to integrate business processes. *Quantitative models* are expected to have a significant impact in the second-wave of reengineering through the deployment of performance and optimization models, economic analysis, and decision support systems. In this paper, we focus on the vital role that quantitative modelling techniques such as those founded in Operations Research and Industrial Engineering can play in reengineering supply chains. These quantitative models can extend the business process reengineering concepts to provide a concurrent reengineering framework for modelling the supply chain processes, identifying reengineering opportunities, evaluating design alternatives, guiding the selection of the best alternative, and deploying tools to implement the design. We illustrate such use by surveying current industrial practice and introducing real world examples based on our practical experience in solving supply chain and reengineering problems.

Keywords.   Business process reengineering; supply chain management; quantitative models.

1. Introduction

The commercial potential of a global market as well as the significant advances in information and communication technologies is driving companies world-wide to reengineer their deployed assets and distributed capabilities to form *manufacturing supply chains* that offer significant business advantages. *A manufacturing supply chain is an integrative approach to managing the inter-related flows of products and information among suppliers, manufacturers, distributors, retailers, and customers.* It consists of cooperative inter-linked processes that transcend production, distribution, and transportation functions and permit the coordination of strategies, tactical plans, and operations across these functions. Supply chains are a major departure from the traditional ways in which businesses have been organized, i.e., vertically integrated along functional lines with significant built-in

Constituting manufacturing supply chains therefore requires major structural changes in the way companies are organized to produce and distribute goods and services. A widely adopted approach, with mixed success, in making these changes is *business process reengineering* (BPR). In BPR, like in manufacturing supply chains, the business process is viewed as a horizontal flow of activities. There is a wide range of opinion on what constitutes BPR. Hammer & Champy (1993) define BPR as a breakaway from old rules, adopting a radical approach to changing business. Hammer and Champy's approach involves four basic stages in BPR, namely, mobilization of a team, diagnosis to identify weaknesses in existing design, redesign to create breakthroughs, and transition to roll out and institutionalize the results. Harrington (1992) on the other extreme views BPR as an incremental process improvement approach. Aikins (1993) regards BPR as a redesign process that leverages the potential of information technology.

Manufacturing supply chains and BPR emerged as independent concepts in the mid-eighties and early nineties respectively and as such are well documented. In this paper, we focus on the vital role that quantitative modelling techniques such as those founded in Operations Research (OR) and Industrial Engineering (IE) can play in reengineering supply chains. These quantitative models can extend the BPR concepts to provide a framework for modelling the supply chain processes, identifying reengineering opportunities, evaluating design alternatives, guiding the selection of the best alternative, and deploying tools to implement the design. We will illustrate this framework by introducing real world examples based on our practical experience in solving supply chain and reengineering problems.

The paper is organized in five sections. Following the introduction, the interest in supply chains and the motivation for modelling supply chains is reviewed in § 2. Section 3 looks at the role and impact of quantitative modelling in reengineering. Practical examples illustrating the use of quantitative models in reengineering are presented in § 4. Finally, § 5 offers some conclusions.

2.   Why model supply chains?

Consider for example a large manufacturing company that caters to a global market and has assets deployed world-wide. A scenario analysis based on a typical consolidated financial summary is shown in figure 1 to illustrate the performance expectations that motivate supply chain management. Supply chains impact the five basic variables that are commonly used in measuring business performance, namely, sales, cost of goods sold, expenses, inventory, and accounts receivable.

As is clear from figure 1, the profitability and efficiency of the enterprise can be improved dramatically if gains can be realized along multiple performance dimensions (modelled by progressively changing the basic variables in the scenario analysis). Supply chains by virtue of their comprehensive enterprise-level focus have the potential to realize such concurrent performance. Therefore, from a business standpoint supply chains merit attention and

Impact Variables	1	2	3	4	5
Sales	1.00	1.00	1.02	1.02	1.02
Cost of Goods Sold	1.00	1.00	1.00	1.00	0.99
Expenses	1.00	1.00	1.00	1.00	0.99
Inventory	1.00	0.90	0.90	0.90	0.90
Accounts Receivable	1.00	1.00	1.00	0.95	0.95
Profitability					
Return on Equity	18.36%	18.36%	24.77%	24.77%	27.64%
Return on Net Assets	15.61%	15.81%	18.81%	19.28%	20.65%
Return on Sales	10.55%	10.55%	12.31%	12.31%	13.18%
Net Profit Margin	5.72%	5.72%	7.57%	7.57%	8.45%
Efficiency					
Inventory Turns	6.57	7.31	7.31	7.31	7.23
A/R turns	3.07	3.07	3.14	3.30	3.30
Total Asset Turnover	0.90	0.90	0.92	0.93	0.93
Net Asset Turnover	1.48	1.50	1.53	1.57	1.57

Notes:
- Scenario 1 represents the base case or the current performance,
- Scenario 2 represents a 10% reduction in inventory costs from the base case,
- Scenario 3 adds a 2% gain in sales revenues over the base case to scenario 2 and so on.
- The profitability and efficiency measures are based on standard financial management definitions.

Figure 1. Scenario analysis illustrating potential gains from supply chains.

capabilities to produce and deliver is equally if not more important than the products themselves. Consequently, in companies that are market leaders, the building blocks of corporate strategy are not products and markets but business processes that constitute supply chains (Stalk *et al* 1992). At the same time reengineering business processes to form supply chains, to realize the performance projected in figure 1, introduces formidable challenges.

The nonlinear impact of the individual performance dimensions on overall profitability and efficiency is readily apparent from the analysis presented in figure 1. In fact, this nonlinearity is inherent in the complex and conflicting trade-off that characterizes the operations of supply chains. The large-scale physical production and distribution network for material flow, the uncertainties associated with the external customer and supplier interfaces, and the myriad cross-functional and nonlinear dynamics associated with internal information flows, are some of the factors that contribute to the supply chain complexity. Lee & Billington (1992) have studied supply chains in electronics, computer, and automobile industries and identified a number of pitfalls that companies face due to the complexity of supply chains. Some of these include the lack of supply chain performance metrics, simplistic inventory policies, ignoring the impact of uncertainties, poor coordination among divisions constituting the supply chain, and inadequate consideration of inventory and response time factors in economic analysis. They note the need to develop an understanding of the supply chain complexities to avoid these pitfalls.

Modelling is fundamental to understanding supply chain complexities and putting the insights to practical use in supporting and guiding business process reengineering. Cypress (1994) has noted that the first-wave of reengineering (during the first half of the nineties) focused on making organizational changes and used primarily *information models of supply*

chains to integrate business processes. In this paper we are interested in the *quantitative models* of supply chains which are likely to have a significant impact in the second-wave of reengineering through the deployment of performance and optimization models, economic analysis, and decision support systems (Slats *et al* 1995). From here on modelling refers to quantitative modelling.

One of the earliest efforts of modelling a supply chain can be traced to Forrester (1958) who developed a simulation model of a production–distribution system to study its time-dependent dynamic behaviour. Another early work by Hanssmann (1959) described an an-alytical model to determine optimal inventory levels in a manufacturing system composed of material procurement, production, and distribution elements. Cohen & Lee (1988) were among the first to propose an analytical modelling framework to evaluate the performance of supply chains spanning functions from raw material procurement to finished goods de-livery to customers. They have applied this framework to supply chains of automobile spare parts and personal computers (Cohen & Lee 1990). Pyke & Cohen (1990) developed a gen-eralized Markov-chain model of a multi-product, three-echelon, production–distribution network and studied its performance.

A recent review of OR models in supply chains (Slats *et al* 1995) notes that most of the quantitative models that are reported in the literature are based on optimization and simulation. They observe that utilization of OR techniques and models to analyse the per-formance of the overall supply chain remain uncommon. The authors conclude that realistic OR models can make considerable contribution in diagnosing and redesigning business processes and facilitate the integration of supply chains. We concur with this observation based on our experience of using OR/IE models in reengineering supply chains.

One such model was for a large electrical component manufacturer (with an annual revenue of $700 M) who operated 10 final assembly plants throughout the US. Some key components were manufactured in-house in two plants in North America. Other raw materials were purchased. The 3000 or so active products were then distributed through a two-echelon warehouse system, shared by several distribution channels (with each channel focusing on a particular market segment). There were also products imported from overseas plants owned by the company. The lower echelon alone had five distribution centres serving the US market, with a small amount of export. This physical supply chain was managed by traditional business functions, such as forecasting, inventory management, production planning, transportation management etc., using a commercial supply chain management system augmented by in-house developed software. This scenario is fairly representative of discrete product manufacturers except for speciality suppliers. The performance model for this supply chain is described in § 4.1.

3. Quantitative modelling and reengineering

BPR in a broad sense has come to signify the following:

- development of a shared understanding of the business processes,
- streamlining and integrating business functions both internal and external to the orga-nization, and

to manage and support planning, sourcing, making, and delivering activities that constitute a supply chain. In practice, companies have taken different tracks to reengineering supply chains that will offer high value at low cost. Some companies have concentrated on making process improvements while others have focused on upgrading and installing information systems. The results on the reengineering activities have been mixed. Gartner Group estimates that nearly 70% of typical BPR efforts fail to achieve the projected results. Although BPR efforts hinge on several factors for success, one vital factor is the availability of tools and an approach that allows BPR practitioners to effectively communicate among themselves as well as with IT professionals who are charged with implementing BPR related systems work. We believe that quantitative models and tools can be very effective in this context.

Despite the high visibility and cost of BPR efforts there is a lack of tools to support the reengineering activities (Childe *et al* 1994). A recent survey of the BPR tool market reveals that there are slightly over 50 or so programs widely ranging in price and functionality from $600 for flowcharting software to $15 K for simulation modelling packages (Barrett 1996; Rovira 1996). These commercially available tools can be grouped into three categories.

(1) Information System departments in organizations have been traditionally using system development tools in BPR. These tools are capable of documenting processes qualitatively in terms of structured functional and input/output diagrams. The oldest and most widely used tool in this category is based on the IDEF (Integrated Definition for Function modelling) methodology that was developed by the US Air Force in the seventies for its integrated computer-aided manufacturing program [**URL 1**]*. Extensions to the IDEF methodology have been developed for enterprise modelling (Malhotra & Jayaraman 1992). An example of commercial implementation of the IDEF methodology is WizdomWorks [**URL 2**].

(2) Groupware and computer-aided software engineering (CASE) vendors are incorporating graphical engines for business process modelling. The groupware tools have the ability to check consistencies in the process and workflow descriptions while the CASE tools provide an object-oriented representation of the business rules and even automatically generate codes for some software implementations. Although these tools are clearly an advancement over IDEF which is based on structured analysis and design they are only capable of qualitatively capturing processes. A list of workflow-based tools for BPR can be seen in [**URL 3**]. An example of workflow based tool is Business Process Modeler [**URL 4**] and an example of CASE based tool is ERwin/ERX for Teamwork [**URL 5**].

(3) At the high-end of BPR tools are discrete event simulation packages that provide tailored BPR modelling capability. These tools go well beyond describing the process and provide a means to quantitatively model supply chains and evaluate performance (Connors *et al* 1995). These tools include BPMAT [**URL 6**] developed by IBM, ARENA [**URL 7**] by Systems Modeling Corporation, and Extend + BPR [**URL 8**] by Imagine That Inc. The use of Extend + BPR is illustrated through the well-known

and Hansen (1994). Other general purpose simulation packages have also advertised BPR modelling capabilities.

A general list of BPR tools can be seen in [**URL 9**]. BPR tools with qualitative process modelling capabilities are inadequate in reengineering supply chains. The process descriptions such as those generated by IDEF are static functional snapshots and are very limited in articulating the existing process. Besides, these tools completely fail to capture the process dynamics and thereby ignore vast amounts of process knowledge. On the other hand, BPR tools with quantitative modelling capabilities can more precisely model processes and provide an effective means for visualizing, benchmarking, and articulating the BPR results. The role of quantitative modelling can best be appreciated when viewed in the context of the BPR approach.

Figure 2 shows the classical BPR approach (Hammer & Champy 1993). The BPR approach has four basic stages: mobilization to form a reengineering team and generate a process map, diagnosis to develop an AS-IS understanding of the processes and identify areas for improvement, redesign to generate process breakthroughs and identify best design from several alternatives, and transition to pilot the TO-BE implementation and eventually roll out the results. These four steps are typically executed in sequence although there are significant iterative dependencies in the diagnosis and redesign stages. One reason for the sequential execution of the diagnosis and redesign stages is that most of the existing BPR tools that can qualitatively document processes do not support concurrency in these BPR stages.

In our view quantitative modelling can have an impact in the diagnosis, redesign, and transition stages of BPR. Further, quantitative modelling can introduce concurrency in the diagnosis and redesign stages as shown in the so-called "concurrent reengineering" approach in figure 3. As indicated at the outset, quantitative modelling can serve as a compelling and useful framework in documenting the supply chain processes, identifying reengineering opportunities, evaluating design alternatives, guiding the selection of the best alternative, and deploying tools to implement the design.

We have encountered in practice many common perceptions among management and staff within a company that have impeded the use of quantitative models. One such perception is that reengineering is a business problem and that quantitative models are not really applicable. Business problems are often founded on complex trade-offs. Quantitative models are particularly suited in identifying the factors that affect the trade-off, ascertaining the assumptions, and objectively framing the issues. Another common perception and an impediment to the use of quantitative models is that they are difficult to build and comprehend and are not process-oriented. With advances in quantitative modelling techniques (heuristics, simulation, and optimization), the availability of a number of rapid modelling tools, and easy access to powerful computing platforms it is possible to build large scale models quickly and animate/visualize them for easier comprehension. This is reinforced by examples in § 4.

4. Use cases of quantitative models in reengineering

In this section we illustrate the use of quantitative models in the diagnosis, redesign,
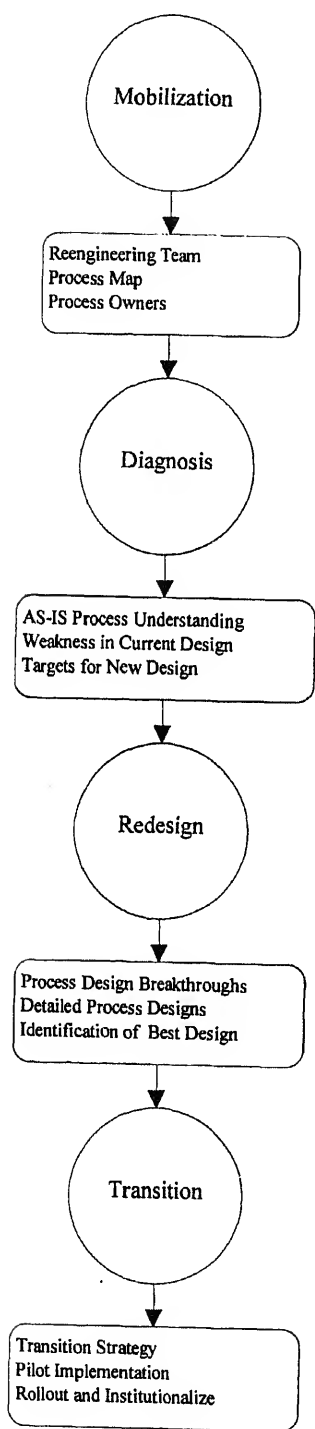
Figure 2. Sketch of classical BPR approach.
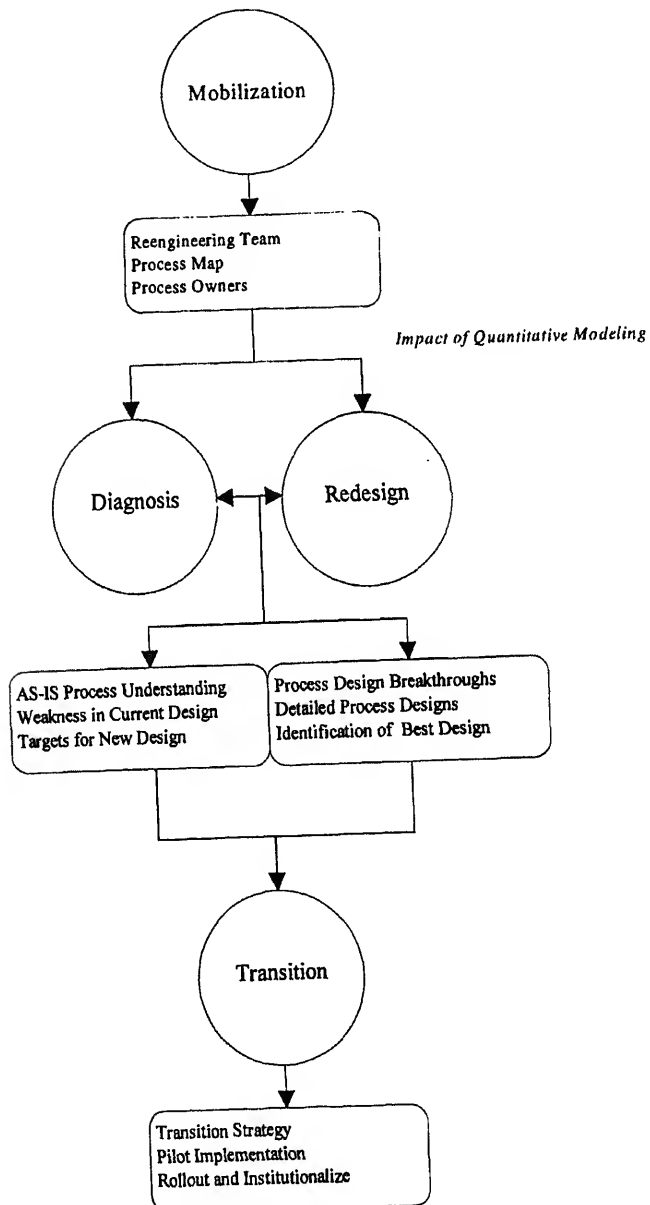
Figure 3. Sketch of concurrent reengineering approach.

ntified above, through some use cases. Section 4.1 samples the current state of industrial
ctice in this area, while §§ 4.2 to 4.4 detail some of our practical experience in applying
antitative models.

Overview of state of industrial practice

pply chains are usually complex, and it may not be obvious from the symptoms where to
gin in terms of rethinking and redesigning the operations. Cypress (1994) observed that
ny reengineering projects are not successful because of the chosen scope of the project
elf. If one had unlimited resources, the ideal approach would be to start from a blank
et and design the business from scratch (Hammer & Champy 1993). More commonly,
ngineering projects have to be carried out with limited resources of budget and time. It
lso very likely that reengineering efforts are initiated when business is not performing
isfactorily, placing severe constraints on the resources available for reengineering. It is
refore prudent to diagnose the supply chain quickly and accurately, so that resources
be targeted to areas that will yield the most improvement in overall performance.

An early example of using quantitative tools for diagnosing production-distribution
tems is described in Hax *et al* (1980). In this study, the production and distribution
nning, sales forecasting, and inventory management systems of a consumer package
ods manufacturer were analyzed. First, a general data analysis was performed to describe
supply chain (such as what fraction of sales is generated where, and what products
contributing to what fraction of sales). An analysis of variance and exploratory data
alysis were performed to analyse the forecast errors. The results give a direction of what
ecasting strategy would be most beneficial (in this case a product-oriented approach, as
oosed to a geography oriented one was picked). A simple, aggregate inventory model
s built to examine the potential savings from reduced inventory levels. This model
owed that safety stock was the largest component of inventory and identified the factors
antitatively.

A more recent example of model-based diagnosis is the study for the electrical com-
nent supply chain as described in § 4.2. A high level quantitative model was developed
investigate the dynamic behavior of the supply chain (Bhaskaran & Leung 1996). This
rformance model was composed of integrated analytical and simulation model compo-
its. A set of representative products were analysed and the goal was to determine high
pact opportunities for an overall supply chain performance improvement. Through a se-
s of what-if analyses, improvement opportunities were identified and ranked according
their impact. An important conclusion of this diagnosis was that improving the forecast
or will not result in the kind of improvement management expected, contrary to popular
lief within the company. Since the model covered all the key components of the supply
ain, we found it useful as a documentation for the (then) current system, as well as a
ol for redesign. This model is described in more detail in § 4.2.

Redesigning a supply chain involves either one or both of the following: redesigning
business processes that constitute the management and operation of the supply chain
d redesigning the physical production–distribution network that executes the material
w in the supply chain. Performing the latter often implies the execution of the former.

views.

A key business process for a manufacturer is order fulfillment. This process defines how a company fills customer orders, starting from the receipt of a customer order and ending with the physical arrival of the ordered products at the customer specified destination. Depending on whether it is an assemble-to-order or assemble-in-advance situation, the order fulfillment process can include the actual final assembly of products. An example of using quantitative models to redesign an order fulfillment process is described in Feigin *et al* (1996). In fact, in this study, a simulation model of the supply chain of the IBM PC Company in Europe was developed to redesign the overall order fulfillment from assemble-in-advance to a combination of assemble-to-order and assemble-in-advance. New inventory planning policies and ways of allocating constrained components were introduced. The simulation model was used to quantify the benefits of these new processes and identify conditions under which they will be most beneficial.

In a similar spirit, simulation models developed with BPMAT, were used in a number of reengineering projects in IBM and others [**URL 6**]. For example, it was used in investigating opportunities for cycle time and cost reduction of a purchasing process. Various staffing and outsourcing alternatives were modelled during the process redesign stage. In the automobile industry, BPMAT models were developed in reengineering a wiring-harness predesign process and an engineering change implementation process. A major goal of both these efforts was to reduce the cycle time involved.

On the supply side, a key process is production planning and scheduling. Fisher *et al* (1994) describe how a fashion-ski-apparel supplier reengineered its supply chain to cope with a changing business environment marked by uncertainty in demand. The company shortened its order processing times, changed the way it purchased raw materials, persuaded their customers to place orders sooner, and adopted a new forecasting and production scheduling approach. In particular, a quantitative model was developed for the purpose of creating a production schedule that minimized the risk of making a non-popular product. The model could also be used to analyse parametric changes to the physical supply chain, such as production capacity. The model was used to guide the continual refinements to the supply chain and product redesign process. It illustrates the concept of "concurrent reengineering" where diagnosis and redesign are done iteratively.

In Feigin *et al* (1996), the authors mention a new production planning process that takes into account constrained availability of components. Instead of performing a traditional MRP type calculation driven by finished goods requirement forecast, the production plan was formulated based on the most desirable allocation of the constrained materials (e.g. for maximizing revenue, fair allocation, product priority). An optimization model was developed to support this decision (Dietrich *et al* 1995).

In design/redesign of physical production–distribution network, the use of quantitative models are better established and a number of commercial model-based tools are available. An early practical work in this area is that of Geoffrion & Graves (1974) (see also Geoffrion 1976). In this study, the problem of selecting sites for distribution centres, their sizes and customer zones, and the transportation flow patterns is modelled by a mixed integer mathematical program. The basic mathematical programming model was used by Hunt-Wesson

ds who were interested in conducting a systematic study of the entire distribution sys-
to investigate opportunities for improvement and to resolve expansion and relocation
es. On the basis of the model recommendations, some existing distribution centres were
cated and new distribution centres were opened, while realizing significant annual cost
ngs. Recently, Pooley (1994) extended the basic model to include production facilities
study for Ault Foods in Canada. Decisions related to production, such as what products
t be produced in which plant and what plant should supply which distribution centre,
e considered simultaneously with distribution decisions.

imilar in spirit, but larger in scale, to the Ault work is the global reengineering project
ied out by Digital Equipment Corporation to restructure its supply chain on a world-
e basis (Arntzen *et al* 1995). Changes in the business environment such as rapid ad-
cement of small computer and communication network technology led Digital "to
vent itself" in order to survive in the new environment. It changed its conventional
tegy of high vertical integration and focused on several core technologies. For exam-
it stopped producing power supplies and cables but retained semiconductors, modules,
systems. At the same time, the overall level of business decreased and the demand
ern had changed from moderate numbers of complex orders for large systems to large
ibers of PCs and workstations. It was clear that Digital's supply chain needed to be
esigned. A large scale quantitative model was developed to aid this process. The model
sidered component supply (which vendors from where), production (which plants mak-
what), and distribution (what customers to serve from where). It minimized a weighted
ibination of total costs and cycle times. A distinctive feature of this model is the consid-
ion of global aspects of the supply chain, such as duties, taxes, trade and local content
iirements. Besides redesigning the supply chain for existing products, the model was
used to configure supply chains for new products. The model was formulated as a
ed integer linear program.

model of a similar flavor was developed at General Motors (GM) earlier, originally
lemented in 1974 (Breitman & Lucas 1987). The original intention of the model was
elp determine the best strategies for placing facilities to support new products in new
kets but was quickly expanded to include decisions such as those faced by Digital.
model was solved by (then) commercial mixed integer programming packages such
IPSX or SCICONIC (Sharda 1995), and was embedded in a decision-support system
business-oriented user interface.

At a more detailed level, the Delco division of GM performed a redesign of its distribution
vork by investigating its shipping strategy (Blumenfeld *et al* 1987). The goal was to
ice the inventory and transportation costs of Delco parts supplied to GM assembly
its. Alternatives included shipping parts from Delco plants directly to GM assembly
its, through a Delco warehouse, a combination of the two, or delivering to several GM
its with one truck load. A quantitative model was developed not only to identify the
imal strategy, which can serve as a benchmark for practical but simpler strategies, but
provide an objective means of analyzing the trade-off between transportation and
entory costs. The optimization model was solved by mathematical decomposition of
distribution network into individual links (Blumenfeld *et al* 1985).

n recent years, there have been a number of commercial software packages devel-

considered by Geoffrion & Graves (1974), and Pooley (1994) as mentioned before. These include PHYDIAS by Bender Management Consultants, Logistics Toolkit by CAPS Logistics [**URL 10**], and Locate 3+ by Cleveland Consulting. Most of these packages are based on solving a large scale mixed integer or linear mathematical program. A comprehensive survey of commercial tools in this area can be found in Leung (1995) and Ballou & Masters (1993). Although not designed for business process reengineering *per se*, they are very useful in reengineering physical supply chain networks. In addition, general purpose operations research/optimization toolboxes or packages can and have been used to solve distribution network design problems. Gangoli & Jenkins (1988) report using SAS [**URL 11**] to improve the distribution network of Warner-Lambert as part of its strategic planning process.

Sometimes it is not necessary to use advanced optimization models for distribution network design. Mercer & Tao (1996) used a relatively simple dynamic model to study alternative distribution policies of a food manufacturer supplying Tesco, a major grocery retailer in the UK. The model helped in deciding how products would be transshipped between the manufacturer's finished goods warehouses to satisfy Tesco's needs.

In the last several years, a notable development related to reengineering supply chains has been the concept of product design for supply chain management (Lee 1993). It is also known as *design for logistics* (Mather 1992). Similar to the well-known concept of *design for manufacture*, design for logistics is concerned with product design that allows effective delivery of products to customers. In particular, a key approach is delayed product differentiation in which the differentiation features within a product family are manufactured into the products as late as possible. For this purpose, an existing product may have to be redesigned or the design of a new product may result in higher material/manufacturing cost. In addition, semi-finished products carried in intermediate stockpiles may now bear a higher per unit value. It is important to analyze whether the resultant savings in supply chain costs (such as reduced finished goods inventory) and other benefits (such as increased flexibility leading to better customer service) can offset the incremental cost. Quantitative models are useful to this end. Lee (1996) presented two inventory performance models to support product design for delayed differentiation, and their real application cases (see also Lee & Billington 1995).

One area that we do not see much reported in the technical literature is the use of quantitative models in the transition phase of a reengineering project. After a new process has been finalized, a roll-out is required to phase-in the new process. A key factor for the success of the new process is the training of staff who either manage or operate the process. Because the nature of reengineering stipulates that the new process may be totally new to the organization, management by experience alone is often not adequate. Here quantitative models of the process developed during the design phase can be valuable in familiarizing the staff with the process and providing management insight on the process dynamics. The latter cannot be gained from static models such as process charts or IDEF diagrams. Simulation models with graphical user interfaces are highly preferable.

Another important aspect of transition is the development or reconfiguration of information systems (transaction processing and decision support) to support the new process. Quantitative models developed during design are now useful as a building block for decision support systems or as a technology enabler of the new process (see § 4.3). Once

model is embedded in a decision support system, it can be used routinely for contin-
improvement of the process. Furthermore, quantitative models are very valuable in
ing operational or tactical planning systems. Typically, an operational system such as
entory/distribution requirements planning has many user specified parameters. While
parameters together make the system very flexible, it is often not obvious how they
uld be set to meet the business objectives. A quantitative model that captures the dynam-
of the system (e.g. the one mentioned in § 4.2) is a tool to help optimize the parameter
ues.

A supply chain performance model

nagement of supply chains that are engaged in the production and distribution of goods
uire a number of key management functions such as demand forecasting, inventory
nning and control, production planning, manufacturing, and distribution. Ideally, these
nagement functions work in cooperation to transform raw materials to finished goods
deploy them appropriately to meet the needs of the market. The presence of a large
duct portfolio, diverse markets, and uncertainties associated with the supply and de-
nd processes are some of the prominent factors that contribute to the complexity of the
ividual management functions as well as their coordination in the overall management
supply chain. Obviously, such complexity tends to make the decision making process
icult and challenging at the strategic and tactical levels of supply chain management.
Managing this complexity requires the visibility of how local decisions, within each
nagement function, affect the overall performance and the competitive advantage of the
ply chain. Supply Chain performance models provide insight on the impact of local
isions on global performance by exercising an aggregate quantitative dynamic model of
supply chain with various decision scenarios. We describe such a quantitative supply
in performance model. The development of the model was part of a larger effort in
roving the profitability and customer service of the business through better strategic
tactical decision making. The model development included: Interviewing decision
kers responsible for strategic and tactical decisions, gathering information on how the
siness operates, understanding the management requirements, and then building and
idating the quantitative model.
The supply chain that was modelled is shown in figure 4. It is a scaled representation of
real supply chain and consists of the following physical elements: Ten representative
ducts, three critical components (i.e., raw materials for final assembly) that are required
the production, a factory comprising two machine groups where the products are
duced, a source warehouse that accepts the factory output, and a distribution system
nprising ten field warehouses located all over the United States to cater to the national
rket. The supply chain comprising the aforementioned elements is managed collectively
the following essential management functions: Demand forecasting, inventory planning
control, production planning, manufacturing, and distribution.
The scaling of the actual system, to derive the supply chain shown in figure 4, was
essary to define a manageable scope for the quantitative modelling exercise. It was
ne in cooperation with the management personnel of the supply chain to ensure that the
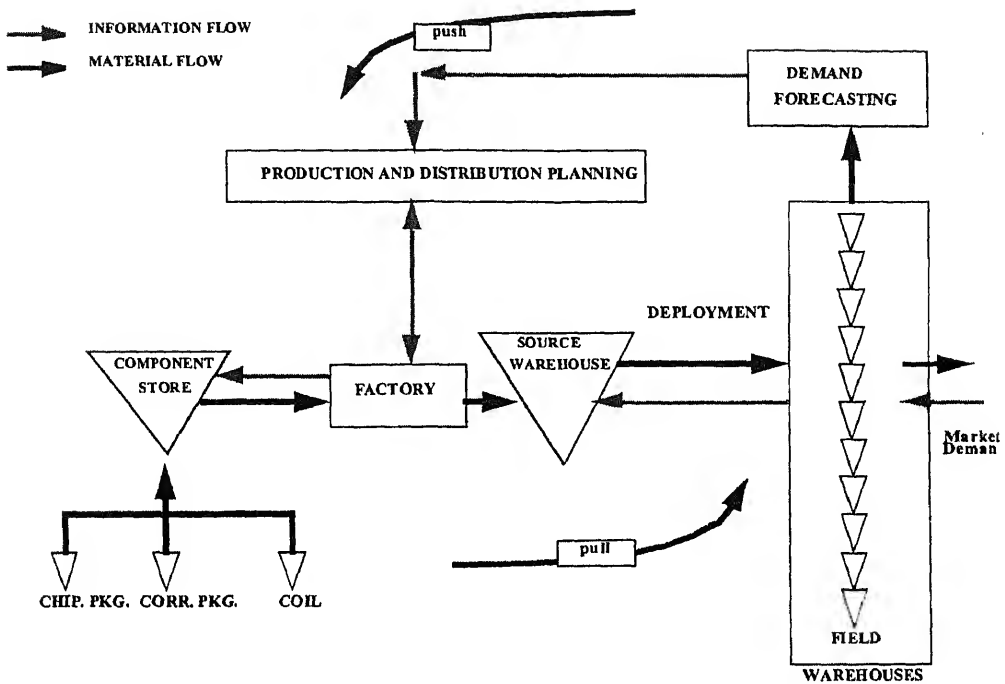
Figure 4. Schematic of supply chain information and material flows.

A detailed study of the overall operations was also conducted to understand the individual management functions and how they collectively worked within the supply chain in the real world.

Operationally, the supply chain can be classified as a make-to-stock business that is forecast driven. Specifically, no customer requests for products are directly entertained at the factory. As a consequence the supply chain can be viewed as a large scale *push–pull system*. The *pull* and *push* parts are referred to as the *commercial* and the *manufacturing* processes respectively. The commercial process is engaged in "selling" products to customers from the inventory that it previously "bought" from manufacturing. The quantity it "buys" at any point in time is governed by the forecast for product demand as well as its policies with respect to hedging against demand and stock replenishment uncertainties. On the other hand, the manufacturing process strives to make-to-stock based on the anticipated needs of the commercial organization. The push and the pull operations are phased apart by the manufacturing lead time, which is the time in advance that manufacturing plans to meet the future needs of the commercial process.

The development of the performance model involved the following steps.

(1) *Defining the supply chain*: This includes the structural constraints of the supply chain model, i.e., what the elements of the supply chain (physical elements such as the source warehouse and the management functions such as production planning) are, and how they operate as part of the supply chain.

(2) *Specifying the model input*: The model input includes the management variables and

model and constitute an independent set of variables (that are different from other internal variables of the model). They are chosen based on the study of the real supply chain and in consultation with management personnel who are the end-users of the performance model. Some of the management variables include the forecast uncertainty for the demand forecasting function, planning frequency and time fence for the production planning function, lead time and delivery reliability for the component supply, run quantities and standard efficiencies for the factory, transit lead times for the distribution, and order quantities and service levels for the inventory planning function.

Specifying the model output: This involves the specification of the performance measures that will be used to measure the performance of the supply chain, upon execution of the model. The supply chain performance is measured using pre-defined performance measures that are categorized as inventory-based and time-based. Inventory-based measures include on-hand inventory throughout the supply chain, inventory turn, and fill rates. Time-based measures include the customer order lead time, the manufacturing lead time and the overall system response time.

Characterizing the elements of the supply chain: Mathematical models of the management functions in the supply chain (shown in figure 4) such as demand forecasting, production planning, and inventory planning, and the physical elements such as the factory, the source warehouse, and the field warehouses are developed. These models are sufficiently detailed to capture the impact of the management variables (item 2) on the performance measures (item 3).

Developing the quantitative model: The individual mathematical representations of the elements in the supply chain are unified in a mathematical framework taking into account the various constraints. Such a quantitative characterization of the supply chain is referred to as the performance model.

It may be noted that the two flows that essentially integrate the supply chain are the material and information flows. The dynamics of these flows are shaped by the external factors such as the customer demand as well as the local decisions made by each management function. The two basic attributes of the dynamics are *quantity* and *time*. Recall that the performance measures for the model are also aptly categorized as inventory-based (i.e. quantity-based) and time-based. Any integrating mechanism must therefore feature these two basic attributes. Additionally such a mechanism must be consistent with the models of the individual elements of the supply chain. The two attributes, namely, the quantity and the time, are characterized stochastically.

The model was validated with real data and presented to decision makers in the form of a decision support system with interactive user interface. The system allowed decision makers to run alternative supply chain scenarios by changing the management variables. These scenarios could then be compared in terms of the performance measures. The system was used in the diagnosis of the overall supply chain as well as in streamlining the existing supply chain operations to improve profitability and customer service.

4.3   *Reengineering the supply chain*

As mentioned in § 2, the supply chain depicted in § 4.2 was managed by function (sales, customer service, forecasting, inventory planning, production planning etc.) and further by market channel (retail, wholesale, industrial, export etc.) on the sales and service side. In such an environment it was not surprising to find the following key observation in the diagnosis phase of an effort to reengineer the order fulfillment process: The process was fragmented with too many hand-offs of both information and physical product, and did not present a consistent, unified interface to the customer. A large part of the redesign was therefore devoted to address this issue.

One candidate design for a new order fulfillment process involved the integration of most of the customer interface and front end planning activities into a new sub-process, called the logistics account management process (LAMP). Specifically, LAMP included taking and delivering customer orders, and was responsible for ensuring that there was adequate stock in the distribution network to fulfill orders. This meant that sales forecasting and finished goods inventory planning would be performed in the same process that dealt with customer orders. Knowledge on customer demands would reside within this process. A new class of jobs, called logistics account managers, had to be created to man the LAMP. The job of logistics account managers was composed of tasks then carried out by several functional departments, and a few new activities which were not performed before. New skills had to be acquired by either training of existing employees or through new hires. A business issue was then how many of such logistics account managers were needed and what customers should be allocated to each team in terms of workload. Would the new process be too expensive (in terms of trade-off between the new supply chain capability and any additional cost)?

In the initial design stage, the principal process flows in the LAMP were captured in a queuing network model using a commercial software package, MPX by Network Dynamics. The objective of the model was to:

(1) establish process feasibility, in terms of staffing levels and customer order response times;

(2) aid the design of customer portfolios, from the view point of work content and load balance of candidate portfolios;

(3) support the development of detailed process design.

A representative customer region was selected for this initial design study. For input to the model, relevant historical demand data were obtained from corporate databases and work flows of existing tasks and estimates of their times were gathered from record files and interviews with appropriate personnel. In establishing process feasibility, quantitative results from the model were provided to the reengineering team and company management to help decide on whether to continue with this design. Some examples are:

(1) minimum total number of logistics account managers required for the customer response desired;

existing marketing organizational structure (i.e., each channel would have its own team of logistics account managers);

number of logistics account managers required if some tasks were to be assigned to a few specialists (this is a mixture of the current process and the new design).

For strategic reasons the existing marketing structure (by channel) would be maintained, with each channel responsible for its own profit and loss. However, to retain the manufacturing economy of scale (important for a low margin business we were in), the product supply would be shared by the different marketing channels. To maximize risk pooling in the new design, finished goods inventory would not be "tagged" for a particular channel until it was used to fill a customer order, even though the channel planned for it through the MP and was independently responsible for their finances and customer service measurements (option 2 immediately above). This created a problem of how to allocate the inventory costs to the different channels. To this end we developed a quantitative inventory measurement model, whose role in the proposed inventory management architecture is shown in figure 5.

First, the demands for all products were analysed and allocated as a "shared" product or a "private" product to a channel. The criterion for allocation was a specified fraction of sales handled by a channel and a threshold total sales volume. All inventory costs of private products would be carried by the individual channels that owned the products, as these costs were (relatively) too small to be further broken down. For the shared products, the inventory measurement model was applied. The model was based on decomposing the total inventory into logical stockpiles according to the reasons they were there, i.e., the stock drivers. Figure 6 shows the logical stockpiles and their measurement principle. At the end of each accounting period, the measurement model would be executed to allocate
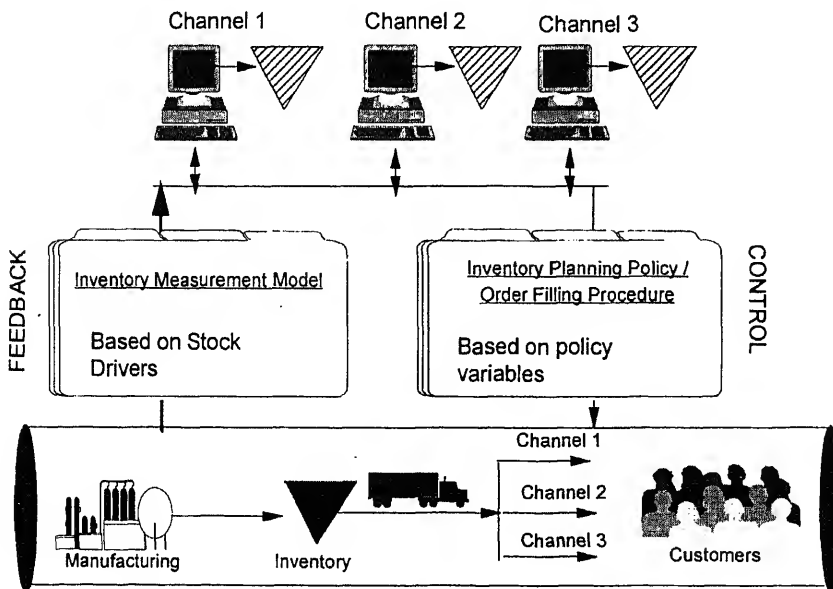
Stock driver	Inventory contribution	Measurement principle
Production Inflexibility	Supply Stocks	Attributed to Supply Process
Distribution Pipeline	Cycle Stocks	Attributed to channels based on Sales Volume
Supply Lead time Variance	Safety Stocks	Attributed to Supply Process
Demand/Forecast Variance at Product Level	Safety Stocks	Attributed to channels based on individual demand/forecast variance
Service Levels	Safety Stocks	Attributed to channels based on target service levels

Figure 6.   Inventory categorization and measurement basis.

inventory costs to the different channels. An important feature of the model was that it was completely transparent to the goods flow and did not impose any burden on the day-to-day operation of the supply chain. The model could be completely automated with required inputs that were already being measured from the physical system. The model served as a technology enabler of the new process design and, more importantly, provided valuable performance feedback to the management of the different channels.

4.4   *A production supply chain*

A single production plant can also be a supply chain with daunting complexities. We consider a North American display product manufacturer with annual revenues of $600 M. The production process is a hybrid of consumer electronics production (in terms of flexible flow through parallel banks of workstations that operate in tandem with issues of major and minor changeovers) and semiconductor production (in terms of batch processing stages, reentrant flows, and associated yield uncertainties). Complicating this scenario was an uncertain demand, vagaries in the supply of components, and a multi-level bill of assembly shown in figure 7 with each level having its own detailed and complex process route through
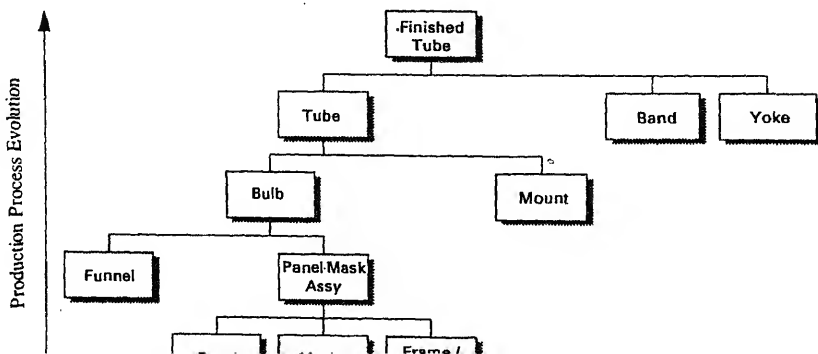
etwork of machine groups. The facility produced three basic families of products.
varying product mix also introduced contentions between these product families for
chine capacities.

he facility was plagued with poor delivery performance and having difficulty coping
the process complexity, machine downtimes, varying demand and product mix, and
ertain component supply. In short, the production planning function had an inadequate
erstanding of the capacity of the plant and was unable to plan effectively. The plant
nagement was interested in reengineering the facility and specifically in evaluating
spective factory layouts. This included outsourcing portions of the overall process and
amlining the process flow in terms of efficient load balancing and process batching.
ddition to these structural and manufacturing execution changes, the management was
eager to upgrade the production planning capabilities of the plant in terms of decision
port tools.

ypically, capacity calculations that accompany production planning are done using
eadsheets which provide a static snapshot of individual machine capacities. The un-
ainty associated with the number of starts, the yield, breakdown of machines, and the
nplicated process flow are difficult to capture in a static rough-cut capacity calculation.
he other extreme, detailed simulation models can capture the process dynamics but are
en too cumbersome to build and too computationally intensive to be of any use as an
ractive decision support tool. An efficient and expedient way to model capacity at an
regate level, that takes into consideration the shopfloor uncertainties and the process
amics, is using *queuing networks* (see figure 8).

Ve used the MPX software to model the process flow based on queuing networks. The
val and departure process to each workstation, the process time at the workstation, the
duction yield, and the workstation reliability are stochastically represented. We then
ve a number of aggregate dynamic equations that collectively characterize the queu-
network. The theoretical basis for the aggregate dynamic models and solution of the
uing network can be seen in Suri *et al* (1993) and Whitt (1983). Optimization models
e also separately built to balance the load across banks of parallel heterogeneous work-
ions with special flow constraints. A decision support platform was built unifying the
X and optimization models and providing an interactive user interface for the analyst.

he input for the aggregate capacity model was obtained from various sources. Process
v data were obtained from process layout drawings and based on tours and actual ob-
vations of the production process. Process times were estimated based on information
vided by production personnel. Process yields were similarly estimated. Machine reli-
lity and changeovers were partly estimated and partly based on historical information.

Material Input   Workstation
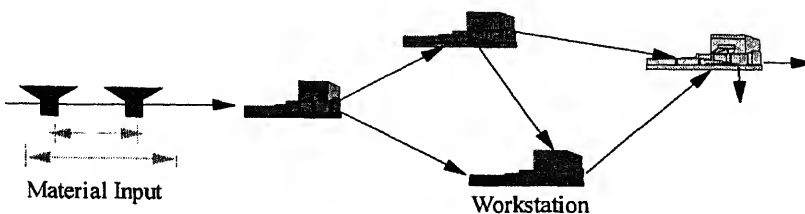
Data on material handling devices such as conveyors and monorails that transported goods between workstations were obtained from technical documentation on operating procedures. Historical demand data were also obtained to characterize the product mix and the arrival process to the queuing network.

The MPX model itself served as a practically useful and comprehensive data dictionary for capacity planning and resource allocation. The model served to integrate the fragmentary data that were stored in various media and validate it in the context of capacity and production planning. Further, the data were now in a form that allowed the factory management to easily evaluate prospective factory layouts as part of its reengineering activities. The specific outputs of the capacity model include long-run average performance for individual products in terms of production achieved, yield, flow time, and work-in-progress. The output also includes utilization for machines as well performance for the overall plant. The production planning personnel envisaged the use of this model for forward (given the demand mix how long it takes to produce) and backward (given the horizon what is the feasible production mix) production planning. Additionally, the model can be used in decision support mode for bottleneck, yield, lead time, and work-in-progress inventory analysis of either the overall plant or any segment of the plant.

5.   Concluding remarks

Quantitative models are very relevant in the diagnosis, redesign, and transition stages of reengineering. They bring concurrency to the diagnosis and redesign stages. Further, the quantitative models provide an objective basis to debate the assumptions and derive alternative designs. Even after the best design is picked the models can be used as part of on-going decision support and planning tools for supply chain management.

Supply chains are of growing interest among businesses as is evident from the rapid growth of supply chain related IT businesses. Following the first-wave of reengineering, organizations have invested heavily in Enterprise Resource Planning (ERP) software such as SAP [URL 12], Baan [URL 13], and BPCS [URL 14] to integrate enterprise-wide transaction data that is contained in heterogeneous databases. At the same time supply chain operational planning software has emerged to offer point solutions such as inventory planning, forecasting, deployment planning and transportation management. In the second-wave of reengineering, organizations are eager to have the various operational planning software inter-operable by integrating them in a flexible manner. Currently no supply chain architecture or supply chain reference model exists that will serve to guide the integration process. At the same time the tactical and strategic planning levels of supply chain management are inadequately served. Quantitative modelling, such as that founded in OR/IE, can serve a useful purpose to fill this void.

However, traditional OR/IE modelling with sole emphasis on models and data structures will not be adequate (Slats *et al* 1995). The models will have to be viewed in the larger context of distributed decision support solutions for supply chain management. The models must therefore be designed to support multiple users to support cross-functional decision making. Additionally, the models must be flexibly designed to be reusable and applied

data and servers hosted in different servers over a network. This requires special ... ...ion to data communication and processing requirements imposed by the model and the ...ution strategy.

...Quantitative modelling, knowledge of business processes that define real supply ...ins, and information technology, collectively constitute a powerful and compelling ...ision technology base for reengineering, integrating, planning, and optimizing supply ...ins.

...ferences

...lications

...ins J 1993 Business process reengineering: Where do knowledge-based systems fit? *IEEE Expert* 8: 2

...tzen B C, Brown G G, Harrison T P, Trafton L L 1995 Global supply chain management at ...Digital Equipment Corporation. *Interfaces* 25: 69–93

...lou R H, Masters J M 1993 Commercial software for locating warehouses and other facilities. *J. Business Logistics* 14: 71–107

...rett R 1996 Chasing the BPR Tool Market. *Enterprise Reengineering* March

...askaran K, Leung Y T 1996 A supply chain performance model. Working Paper, IBM T J Watson Research Center, Yorktown Heights, NY

...menfeld D E, Burns L D, Diltz J D, Daganzo C F 1985 Analyzing tradeoffs between trans-...portation, inventory, and production costs on freight networks. *Transportation Res.* B19: 361–380

...menfeld D E, Burns L D, Daganzo C F, Frick M C, Hall R W 1987 Reducing logistics costs at General Motors. *Interfaces* 17: 26–47

...eitman R L, Lucas J M 1987 PLANETS: A modelling system for business planning. *Interfaces* 17: 94–106

...ilde S J, Maull R S, Bennett J 1994 Frameworks for understanding business process re-...engineering. *Int. J. Oper. Product. Manage.* 14: 22–34

...hen M A, Lee H L 1988 Strategic analysis of integrated production-distribution systems: Model and methods. *Oper. Res.* 36: 216–288

...hen M A, Lee H L 1990 The management of integrated production distribution systems. *Proc. NSF Design and Manufacturing Systems Conference*, Tempe, AZ, pp 259–263

...nnors D, An C, Buckley S, Feigin G, Jayaraman R, Levas A, Nayak N, Petrakian R, Srinivasan R 1995 Dynamic modelling for business process reengineering. IBM Research Report RC 19944, IBM T J Watson Research Center, Yorktown Heights, NY

...press H L 1994 MS/OR imperative: Make second generation of business process improvement mode work. *OR/MS Today* 21(1): 18–29

...etrich B, Connors D, Ervolina T, Fasano J P, Lin G, Srinivasan R, Wittrock R, Jayaraman R 1995 Production and procurement planning under resource availability constraints and demand variability. Research Report RC-19948, IBM T.J. Watson Research Center, Yorktown Heights, NY

...gin G, An C, Connors D, Crawford I 1996 Shape up, ship out. *OR/MS Today* April: 24–30

Fisher M L, Hammond J H, Obermeyer W R, Raman A 1994 Making supply meet demand in an uncertain world. *Harvard Business Review* May–June: 83–93

Forrester J W 1958 Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Rev.* July–August: 37–66

Gangoli N, Jenkins R T 1988 Distribution system design with SAS/OR and the matrix generator. *Proceedings of the 13th Annual SAS Users Group International Conference*, Orlando, FL, pp 194–198

Geoffrion A M 1976 Better distribution planning with computer models. *Harvard Business Rev.* July–August: 92–99

Geoffrion A M, Graves G W 1974 Multi-commodity distribution system design by Benders decomposition. *Manage. Sci.* 20: 50–67

Hammer M, Champy J 1993 *Reengineering the corporation* (New York, NY: Harper Business)

Hansen G 1994 A complex process: The case for automated assistance in business process re-engineering. *OR/MS Today* August: 34–41

Hanssmann F 1959 Optimal inventory location and control in production and distribution network. *Oper. Res.* 7: 483–498

Harrington H J 1992 *Business process improvement* (New York, NY: McGraw-Hill)

Hax A C, Majluf N, Pendrock M 1980 Diagnostic analysis of a production and distribution system. *Manage. Sci.* 26: 871–889

Lee H L 1993 Design for supply chain management: Methods and examples. *Perspectives in operations management: Essays in Honor of Elwood S Buffa* (ed.) R Sarin (Boston, MA: Kluwer Academic) pp 45–65

Lee H L 1996 Effective inventory and service management through product and process redesign. *Oper. Res.* 44: 151–159

Lee H L, Billington C 1992 Managing supply chain inventory: Pitfalls and opportunities. *Sloan Manage. Rev.* Spring: 65–73

Lee H L, Billington C 1995 The evolution of supply-chain-management models and practice at Hewlett-Packard. *Interfaces* 25: 42–63

Leung Y T 1995 Commercial software for distribution network optimization: A survey of current products. White Paper, IBM T.J. Watson Research Center, Yorktown Heights, NY

Malhotra R, Jayaraman S 1992 An integrated framework for enterprise modeling. *J. Manuf. Syst.* 11: 426–441

Mather H 1992 Design for logistics (DFL) – The next challenge for designers. *Product. Inventory Manage. J.* (First Quarter): 7–10

Mercer A, Tao X 1996 Alternative inventory and distribution policies of a food manufacturer. *J. Oper. Res. Soc.* 47: 755–765

Pooley J 1994 Integrated production and distribution facility planning at Ault Foods. *Interfaces* 24: 113–121

Pyke D F, Cohen M A 1990 Effects of flexibility through set-up time reduction and expediting on integrated production-distribution systems. *IEEE Trans. Robotics Autom.* 6: 609–620

Rovira M 1996 Private communication. IBM T J Watson Research Center, Yorktown Heights, NY

Sharda R 1995 Linear programming solver software for personal computers: 1995 report. *OR/MS Today* 22(5): 49–51

Slats P A, Bhola B, Evers J J M, Dijkhuizen G 1995 Logistic chain modelling. *Eur. J. Oper. Res.* 87: 1–20

Stalk G, Evans P, Shulman L E 1992 Competing on capabilities: The new rules of corporate strategy. *Harvard Business Rev.* March–April: 57–69

R, Sanders J L, Kamath M 1993 Performance evaluation of production networks. *Handbooks Operations Research and Management Science* (eds) S C Graves, A H G Rinnooy Kan, H Zipkins (Amsterdam: North Holland) 4: 199–286

tt W 1983 The queueing network analyzer. *Bell. Syst. Tech. J.* 62: 2779–2815

ld-Wide Web References

RL 1] http://nemo.ncsl.nist.gov/ standsp/stand.html; IDEF Standards

RL 2] http://www.wizdom.com/bpr/cpr.html; WizdomWorks Home Page.

RL 3] http://www.ctt.fi/tte/staff/ojp/workflow.html; Workflow related internet resources.

RL 4] http://www.software.ibm.com/ad/promodel/bmtn0mst.htm; IBM Business Process Modeler.

RL 5] http://www.cayennesoft.com/products/erwinteam.htm; IDEF with CASE functionality.

RL 6] http://www.research.ibm.com/sas.html; BPMAT – IBM Software Application and Services.

RL 7] http://www.sm.com/arena.htm; ARENA by Systems Modeling Corporation.

RL8] http://www.imaginethatinc.com/home.html; Extend Simulation Software Home Page.

RL 9] http://duticai.twi.tudelft.nl/lists/bpr-l/faq/tools.html; Archive of tools for BPR.

L 10] http://www.caps.com; CAPS Logistics Toolkit home page.

L 11] http://www.sas.com; SAS Institute.

L 12] http://www.sap.com; SAP AG Inc.

L 13] http://www.baan.com; The Baan Company.

L 14] http://www.ssax.com; System Software Associates Inc.

egrated product development

TAY ENG HOCK

National University of Singapore, Department of Mechanical Engineering,
10 Kent Ridge Crescent, Singapore 119260
e-mail: mpetayeh@nus.edu.sg

Abstract. Manufacturing and design are very closely related. The manufacturing capabilities available impact the scope of design, while design for manufacturing ensures the economic success of the products.

Furthermore, the major goals of firms in the nineties are to significantly reduce product costs and time to market (TTM). To meet these goals, quality product designs that meet customers' needs have to be developed.

Design methods such as Quality Function Deployment and Pugh's concept selection technique have been used to significantly improve engineering design processes. Developed as separate tools, however, they are difficult to integrate and coordinate in the total design process, since the relationship of the two methods is often unclear.

This paper demonstrates, through a case study, that these methods are in fact results of a similar underlying concept. Design problems and solutions are unified by the concept of engineering models. An engineering model is a set of equations that relates the design variables to the performance metrics used to quantify performance of a product. Together with the engineering models, Quality Function Deployment and Pugh's concept selection technique have been used in the design and development of a hematology machine from concept to prototype.

Keywords. Integrated product development; quality function deployment; concept selection technique; time to market; product costs; hematology machine.

Introduction

relationship between manufacturing and design has been well recognized. A study e by Westinghouse (1984) shows that 80% of all life-cycle costs are fixed during the gn phase. Good manufacturing techniques cannot compensate for poor design.

urther, in this highly competitive age, for a company the ability to identify products the customers need is crucial. The degree to which a product satisfies customer desires critical product success factor (Hise *et al* 1990). A consensus is rapidly developing

in industrial practice that customer desires can only be obtained through actual contact with the customer and that designers are often wrong when they try to guess what the customer wants (Rabino & Muskowitz 1984). To facilitate customer focus, several structured methodologies and representations for organizing and presenting customer information have been developed. One such representation is the House of Quality (HOQ), which helps product designers to explicitly identify customer requirements, relate them to objective engineering characteristics, identify tradeoffs, and to evaluate the characteristics of a potential product relative to competing products (Hauser & Clausing 1988). The HOQ provides a product development team with a compact description of three important items

(1) Customer needs and their relation to objective engineering characteristics.

(2) Comparisons with competing products based on objective engineering numbers.

(3) A summary of the engineering tradeoffs inherent in the design.

In a typical situation, marketing staff collect the data about customers and competing products and, possibly with some input from engineering, create the HOQ. They decide a set of performance targets which are then communicated to the designers. There are two specific sources of problems in this process,

(1) Targets set based on the information contained in the HOQ alone are often unrealistic. As a result, designers cannot achieve them and this results in time-consuming iterations until an achievable specification is reached.

(2) The way coupling between design variables is described in the HOQ does not adequately reflect the trade-off that must be made in real design problems.

After understanding the voice of the customers, concepts are generated using brainstorming and these concepts are selected using the method proposed by Pugh (1991). This will be described in a later section using a case study.

In addition to House of Quality and Pugh's Selection Method, designers often have reliable engineering models that they can use to test the limits of product performance. An engineering model is a set of equations that relates the design variables to the performance metrics used to quantify performance of a product.

The synthesis of the methods mentioned above is shown in figure 1.

2. Quality function deployment. The house of quality

The identification of customer needs remains a key aspect of developing robust and successful products. Without customer needs to guide new product design decisions, the products themselves have relatively low probabilities of success in the market. The three key steps that allow a researcher to construct the voice of the customer are as below (figure 2).

(1) Identify unstructured list of customer needs.

(2) Set development priorities by quantifying the relative importance of the different needs.

(3) Translate the quantified needs into specific engineering goals.
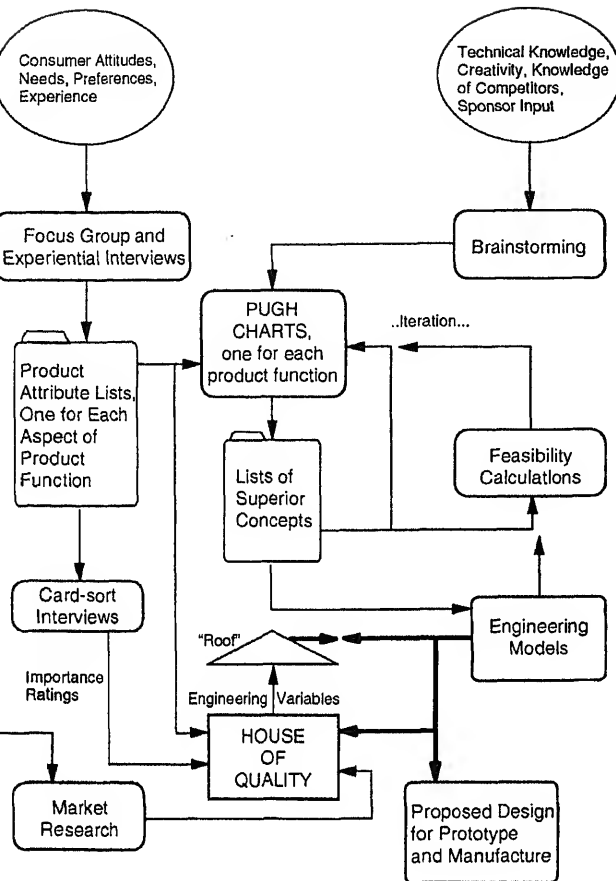
Figure 1. Integrate product design process.

Further, it remains critical that the exact words and vernacular be used throughout the identification, structuring and prioritization process. Without using the exact words of the customer, it may become difficult to identify the nuances of the customer preferences; if the customer voice is constantly processed and re-translated, the final translation may become markedly different from the customer's initial intentions. Because of this need to capture the exact words of the customer, one-on-one interviews and focus groups provide a superior medium other than sales force or distributor reports for producing a rich set of needs.

Further, the interactive one-on-one interviews may provide a more robust set of needs than focus groups because of the opportunity to probe deeper for customer needs, rather than focusing on the obvious issues which a focus group can easily identify.

Hauser and Griffin (Griffin 1989) have also carried out research on identifying how many customers need to be interviewed in order to identify a percentage of the unmet needs. They have created a Beta-binomial model which predicts the percentage of the needs identified given the number of customers interviewed. Based on the model predictions, interviews with sixteen customers identify 80% of the needs, while thirty interviews would be required to identify 90% of the needs. Considering the high cost of market research, significant cost savings can be realized by reducing the number of surveyed or interviewed customers.

Process Steps   **Process Tools**

| Develop Understanding of the Customer in this New Market Segment | → | • Experiential Interviews with Med Techs
• Patent Searches |

| Identify the List of Customer Needs | → | • Experiential Interviews with Med Techs
• Focus Group with Med Techs *(large physician practices)* |

| Quantify the Importance of Each Customer Need | → | • Card-sort Interviews with Med Techs |

| Translate the Quantified Needs into Specific Engineering Goals | → | • House of Quality with Research Team |

Figure 2. Market research process – identifying and applying the voice of the customer.

Hauser and Griffin's model was used to determine the number of customers that we should interview. Through experiential interviews, card-sort interviews, and the focus group, the research team has directly contacted 33 medical technicians who were involved in hematology testing. Based on Hauser and Griffin's model, we should have identified over 90% of customer needs. Some of the results are shown in tables 1–3. Due to proprietary considerations, not too many details are given on this project in this paper. However, the nature of the project is to design a next generation of hematology equipment for a major corporation in the United States of America.

Table 1. Ten most important needs for physician-practice and medical technicians.

Importance	Customer need
1	Achieves level of accuracy equivalent to competitor's machine
2	Accurate even if the blood clots
3	No opening of sample tubes
4	Machine that can be stopped for RUSH job or emergency job ('STAT').
5	Low sample volumes for automatic testing
6	Analysis time less than a minute
7	Eliminate opening tubes to solve aerosol problem
8	Performs repeat test automatically

le 2. Examples of translations from customer needs to engineering design requirements.

tomer needs	Engineering design requirements
chine covers that are easy to remove	• Time to remove cover • Number of steps to take apart for repair • Height, width and depth of the machine
ninate error in patient identification	• Number of human failure points • Amount of information on tube • Level of redundancy • Labeling time • Patient identification system cost • Read and process time for patient identification
re quality control information	• Data storage capacity • Number of computing resources and functions used • Number of external access routes
ninate weekly or long-term maintenance	• Weekly cleaning time • Mean time for preventive maintenance • Downtime for preventive maintenance
Hour service available	• Repair time • 24-Hour maintenance contract availability • Mean time between failure • Lead time for repair
alyse blood in less than a minute	• Centrifuge spin time • Scanning time • Analysis time • Material handling time • Read and process time for patient identification • (result) Print out time

This QFD process is most effective when collective experiences of a multifunctional m are greater than the sum of the individual experiences. Therefore, the decisions de jointly by the team will consider all the important issues of customer needs, robust hnology, manufacturability, and external design.

The relationship matrix in the House of Quality (Hauser & Clausing 1988) helps the m members to clarify the relationships between requirements and output and serves as a erence document as the design process evolves. The roof of the HOQ provides informa-n on the relationships among the engineering characteristics. Additional columns and vs on the HOQ provide summary information on the technical difficulty of improving gineering characteristics, and the relative position of competitors.

Pugh's concept selection technique

e Pugh's concept selection technique provides a systematic process for choosing the est" design out of several, based on attributes that are previously determined. The Pugh pcess guards against potential weaknesses that might be due to lack of thoroughness conceptual selection. The process forces the team to consider all types of possible ncepts, explain the potential solutions in sketch form, and place the solutions along horizontal axis of the selection matrix: the team then identifies relevant criteria for

Table 3. Most important design requirements in physician practices and hospitals (10 for most impor
and 1 for least important).

Engineering design requirements	Physicians	Hosp
Increase variability of batch sizes	4	4
Reduce number of steps before walkaway	6	6
Reduce number of steps on return	4	4
Reduce required time between loading	4	4
Include password for supervisor	5	6
Include alternate technology	5	6
Reduce time for maintenance	5	5
Increase automatic maintenance level	4	4
Reduce number of human failure points	6	7
Reduce material handling time	4	4
Reduce downtime for preventive maintenance	4	5
Reduce number of blood exposures	5	4
Reduce number of aerosols	5	4
Reduce number of times disposables handled	4	4
Reduce number of sharps handled	4	4
Increase data storage capacity	5	6
Reduce number of computing resources and functions used	5	6

the concept selection which could come from either customer input or expectations of
design team.

A datum concept is then selected with which all other concepts will be compared aga
(often, the datum concept is the currently existing design). After establishing the da
for each individual criteria the other concepts are rated as better ($+$), worse ($-$) or
same (S) as the datum. After completing the matrix, it is possible to graphically v
which concepts best satisfy a wide range of criteria. Based on the matrix, it is poss
that a dominant concept may emerge; it is also possible that some concepts will ha
number of relative strengths and weaknesses. Hybridization of the concepts can be pl
alongside the initial concepts and "re-pughed" with the same criteria. Thus, this pro
provides a systematic method of acquiring:

(1) greater insight into specification requirements;

(2) greater understanding of the problem;

(3) greater understanding of potential solutions;

(4) knowledge of why one concept may be stronger than another;

(5) team consensus on all of the above;

(6) a natural stimulus to produce additional concepts.

Moreover, in addition to improving communication across different members in a pro
team, QFD provides a format for documenting decisions that can be used to guide fu
decisions. All of the documents produced by the QFD process can be used as documenta
for design decisions and product modifications. As a design project progresses, the
often confusion, and second-guessing as to why a previous decision was made. The F
and Pugh selection matrices can immediately provide information on the rationale fo
previous decisions. Figure 3 shows the results of one of the Pugh matrices for pa
identification.

Patient Identification Concept Selection Criteria	"Shipping Tag"	"Bracelet & Tube"	"Form & Tube Sticker"	"Form & Tube"	"Existing"	"Auto"	"Coded Tube & Form"
System Cost (-)	+	S		+	+	-	+
Manual human interaction with tube and/or form (-)	-	+		S	-	+	S
Number of potential failure points from human interaction (-)	S	+	D	S	-	+	S
Cost per sample - disposable cost (-)	+	+	A	S	+	+	S
Machine can read patient identification (+)	-	S	T	-	-	S	S
Ease of human reading on tube (+)	+	S	U	S	-	+	S
Reliability in reading patient identification (+)	S	S	m	S	-	+	S
Potential patient interference with blood band and measurement operation (-)	-	+		+	+	+	S
Laboratory information system (LIS) dependency (-)	+	S		+	+	+	+
SUM	**0**	**+4**	**0**	**+2**	**-2**	**+6**	**+3**

Figure 3. Patient identification Pugh chart.

Engineering models

...ny firms designing products, ranging from bearings to automobiles, have developed ...ineering models for their products. An engineering model is a set of equations that ...tes the design variables to the performance metrics used to quantify the performance ... product.

...hese models are used to decide whether designs are feasible, to explore the perfor-...ce envelope of a design without actually building a physical prototype, and to study ...tradeoffs involved in the design. A reliable engineering model is often a key part of ...competitive advantage in product design. This is because it allows the designer to ...eriment with large numbers of possibilities without increasing the design time. Hav-
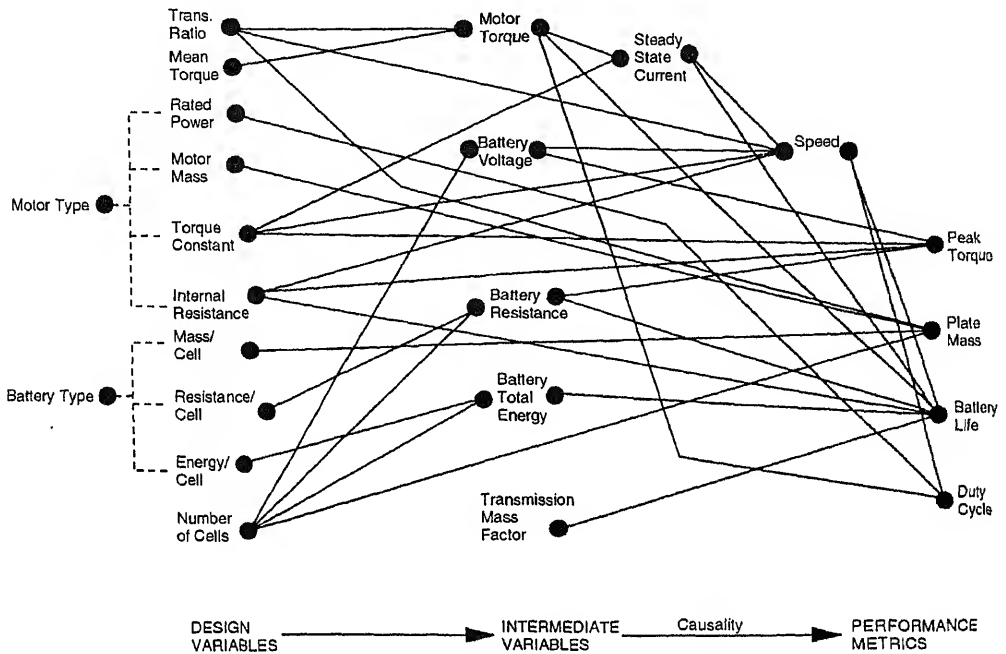
Figure 4.   Structure of engineering model for precision centrifuge.

completed rapidly and reliably on the computer without the delays and effort required to build and test prototype hardware.

Figure 4 shows the structure of the engineering model for the precision centrifuge that is required in the design of the hematology equipment.

The inputs to the model are:

(1) Motor torque constant (K_T), internal resistance (R_M) and mass (M_{mot}), rated load (P_R). All these get fixed when a particular motor type is chosen;

(2) Battery resistance/cell (R_{cell}), battery energy/cell (E_{cell}) and mass/cell (M_{cell}). All get fixed when a particular battery type is chosen;

(3) Number of cells in the battery (N_C);

(4) Transmission ratio (N_T);

(5) Mean torque (τ_M) at which the tool will be operating.

The outputs of the model are:

(1) Weight of the plate (M);

(2) Speed at rated torque (ω_R);

engineering model for the precision centrifuge is therefore:

$$\tau = \tau_M/N_T,$$
$$V = N_C \times 1.2,$$
$$E = N_C \times E_{\text{cell}},$$
$$R_B = N_C \times R_{\text{cell}},$$
$$I_A = \tau/K_T,$$
$$\omega_R = N_T(V - I_A(R_M + R_B))/K_T,$$
$$\tau_{\max} = VK_T/(R_M + R_B),$$
$$L_{\text{av}} = E/(I_A^2 R_B + \tau\omega_R),$$
$$M = M_{\text{mot}} + N_C M_{\text{cell}} + M_t N_T,$$
$$D = P_R/(W_R\tau),$$

ere

τ:	motor torque,
V:	battery voltage,
E:	battery total energy,
I_A:	steady state current,
P:	motor power output,
M_t:	transmissive mass factor.

ome experimentation with the model makes it clear that many trade-offs are possible
he design. For example:

nergy can be added by using more cells in the battery, but this makes the centrifuge
eavier;

more efficient motor can be chosen but efficient motors tend to run at higher speeds
nd so a greater transmission ratio is needed to get the required torque.

ingineering models are useful in conjunction with the House of Quality because they
nplete the chain from the entities the designers can actually change, the design variables,
vhat the designer wants to affect, namely, the customer perception of the product.

Final system configuration

 final system design has two centrifuges, each of which can be loaded with a carousel
ny time. When the centrifuges are expeditiously loaded with carousels batched with 20
es, the machine can have a maximum achievable throughput (after the first 5 minutes
entrifuging) of 400 assays per hour.

 black and white LCD screen displays patients' blood analyses, quality control and
bration information. The system has the following features:

valkaway capability;

liffraction grating optical reading station;

utomatic patient identification;

Laboratory Information System interface capability

6.   Conclusion

The product development team's group dynamics is critical to the success of a product (Finger & Dixon 1989). It is important that the team members interact and communicate effectively and understand each other's discipline. The integrated product development incorporating tools from Quality Function Deployment, Pugh Concept Selection Technique and Engineering Model have been found to be effective in this project which involves people from diverse backgrounds.

References

Finger S, Dixon J R 1989 A review of research in mechanical engineering design. Part II: Representations, analysis and design for the life cycle. *Res. Eng. Design* 1: 121–137

Griffin A 1989 *Functionally integrated new product development*. Ph D dissertation, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA

Hauser J R, Clausing D 1988 The House of Quality. *Harvard Business Review* May–June: 63–73

Hise R T, O'Neal L, Parasuraman A, McNeal J U 1990 Marketing/R&D interaction in new product development: Implications for new product success rates. *J. Prod. Innovation Manage.* 7: 142–155

Pugh S 1991 *Total design: Integrated methods for successful product design* (Addison-Wesley: New York)

Rabino S, Moskowitz H 1984 Detecting buyer preference to guide product development. *J. Prod. Innovation Manage.* 1: 140–150

Sriram D, Logcher R, Fukuda S 1989 An object-oriented framework for collaborative engineering design. In *Computer-aided cooperative product development* (MIT, Cambridge, MA: Springer-Verlag)

Westinghouse 1984 Report on Life Cycle Costs, Westinghouse Productivity and Quality Center, Pittsburgh, PA

lume modelling for emerging manufacturing technologies

V CHANDRU and S MANOHAR

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India
e-mail: [chandru,manohar]@csa.iisc.ernet.in

Abstract. The next generation manufacturing technologies will draw on new developments in geometric modelling. Based on a comprehensive analysis of the desiderata of next generation geometric modellers, we present a critical review of the major modelling paradigms, namely, CSG, B-Rep, non-manifold, and voxel models. We present arguments to support the view that voxel-based modellers have attributes that make it the representation scheme of choice in meeting the emerging requirements of geometric modelling.

Introduction and motivation

confluence of three major technological revolutions is providing the impetus to the antum changes in manufacturing technologies as we move into the next millennium.

Information technology: Increased power of processors, decrease in the cost of memory, evolution and acceptance of the Internet and its rapid growth, the World Wide Web, multimedia and virtual reality technologies, distributed and parallel computing, growth of terrestrial, satellite, and broadcast communications, are some of the components of information technology that have a direct impact on manufacturing. For instance, the information superhighway with teleconferencing facilities will lead to large scale telecommuting of information workers. Computer-supported collaborative work will enable a geographically separated workforce to work as a team.

Material technology: Several new materials called smart materials, materials with memory, composites with highly specialised functionality, low temperature superconducting materials, materials for fibre-optic communication, high efficiency solar energy converters, materials for solid-state batteries etc. have evolved.

Process technology: Advances in VLSI technology have enabled the production of multi-million transistor chips and 256 Mbit memories. The experience with VLSI technology has inspired two diametrically opposite process technologies, namely micro

(also referred to as rapid prototyping technologies, RPT). Both have their genesis in VLSI fabrication technologies and both require highly interdisciplinary research and development. Similar to VLSI technology, these two technologies depend critically on the development of new and powerful CAD tools.

(i) MEMS: In this technology, the sensors, the processors and the effectors can all be fabricated in the same process. An example is a commercially available radar, costing $5 that can be fitted into automobiles to detect potential collisions with other automobiles. MEM systems promise to be the ultimate in mass production of electro-mechanical components, finished assemblies in millions of units at per unit costs of a few dollars.

(ii) RPT: A range of new processes variously called layered manufacturing, additive manufacturing and three-dimensional printing have emerged during the past few years (Burns 1993). These processes fabricate a component directly from the CAD description without the need for any additional fixtures, dies or tools. As opposed to the traditional approach of removing unwanted material from a block to arrive at the desired shape, these new processes build an object by selectively adding material where required by the geometric model. This technology promises to be the ultimate in customization, enabling the economical fabrication of one-of-a-kind finished assembly of complicated subunits.

In addition to the three technological streams the changing economic scenario is forcing a relook at manufacturing strategies. The rapid globalisation of the economy and the increasing competitiveness in the manufacturing sector has led to new paradigms like concurrent engineering, just-in-time manufacturing and business process reengineering. Product cycle times have come down from a few years to a few months and less. From concept to the customer, time required has been constantly shrinking and now we have reached a stage where custom products and low-volume high-custom products have begun to dictate competitiveness. Agile manufacturing is the term coined to reflect this ability to meet ever-changing product specifications. In areas like wearable computing (Finger *et al* 1996), where the associated technologies are constantly changing, completely new design philosophies have become essential.

The combined impact of these changes can be expected to drastically change manufacturing. Some of the extrapolated possibilities are conjectured below.

• Customer-in-the-loop manufacturing: For example, custom kitchens are designed by the customer using VR interfaces, built and delivered later at the customer's residence (Nomura *et al* 1992).

• Coin-operated manufacturing/shopping-mall manufacturing: The previous scenario refers to consumer durables. A similar quest for customization of consumer goods like shoes and personal accessories may result in compact manufacturing units being located in shopping malls, where the customer specifies the size (by placing the foot in a shoe

be used. Examples are: kitchenware (spoons, forks, plates etc.) and toys. The customer uses highly interactive tools to design the objects and fabricates them in-house. The equipment has to be capable of making a range of shapes using a range of input raw material (fresh or recycled).

Space-based manufacturing : As humans expand into space, by first building space stations and later moving to the planets, newer manufacturing technologies will be needed: flexible systems that occupy minimal volume and have minimum weight and are capable of fabricating a range of complex shapes will be required.

Distributed product life-cycle management: Products will be conceived, designed, fabricated, marketed, maintained and upgraded by collaborative teams distributed geographically.

When we look at this exciting spectrum of possibilities, as well as the as-yet unthought of new processes and materials, it becomes very important to examine the tools we need to make them happen. At the fundamental level, any material object (mechanical, electromechanical) to be fabricated needs a model that forms the canvas on which the design process generates diverse paintings. Any such object is captured as a volume model. We use the term volume model since it is more general than the term solid model: a volume model can represent non-homogeneous objects (composites), flexible objects, and objects that are made of solid, liquid and amorphous material.

In this paper we examine the fundamentals of volume modelling for emerging manufacturing technologies. We first outline a set of requirements that a volume modelling paradigm should satisfy. This is followed in § 2 by a brief outline of existing modelling paradigms and their comparison based on our desiderata. In the final three sections we look in turn at the requirements of three major aspects of manufacturing, namely, virtual prototyping, physical prototyping and reverse engineering, and indicate how voxel modelling is uniquely suited to meet these requirements.

The major thrust of this paper is that voxel-based modelling has substantial unexploited advantages over other modelling paradigms.

Desiderata for a volume model

We present below a set of attributes that can provide a comprehensive characterization of a representation scheme for volume modelling. We have grouped these attributes to reflect the three phases in the lifecycle of mechanical parts:

Model creation and maintenance

- renderability
- morphological dexterity
- heterogeneity
- editability
- brevity

- Model analysis
    - analysability
- Model fabrication
    - reconstructibility
    - physical realizability
    - accuracy

1.2 *Renderability*

This captures the ease with which the volume model can be rendered. Parameters to be considered are the complexity of the rendering computation and the quality of the rendered images as well as the ability to render the model at various levels of quality. For interactive modelling, the ability to rapidly preview the model, interact dynamically with the model as well as get high quality presentation images, are all essential.

1.3 *Morphological dexterity*

The ability to represent a variety of shapes is an important aspect of any volume modeller. Traditionally solid modelling has focused on representing shapes that are restricted to manifold topologies. However, the importance of non-manifold representations has been recognized (XOX Corporation 1995).

The morphological dexterity attribute is aimed at capturing the complexity of shapes that can be represented and the ease with which complex shape modifications can be achieved on the given volume representation.

1.4 *Heterogeneity*

Traditional CAD tools available today implicitly assume that the part being designed is to be fabricated using a single homogeneous material. A few niche areas like aircraft/spacecraft use specialised CAD tools to model composites. The need for general CAD tools to handle composites will be realized as an increasing range of products are fabricated with composites.

The heterogeneity attribute will evaluate the ability of a representation to capture inhomogeneities of the following types: use of different materials, use of different densities of the same material across the cross-section of the solid, use of hollow structures (example: honeycomb) and integrated electromechanical components.

1.5 *Accuracy*

When evaluating a volume representation scheme with reference to the accuracy of the representation the following factors have to be considered:

The accuracy with which the model captures the intended shape of the part is of primary importance. Of equal importance is the ease and accuracy of specification of tolerances consistent with the ideal design (Requicha 1984).

Editability

e central purpose of a computer-based modelling paradigm is to empower the user to
ect changes to the model. We intend the broad term "editability" to encompass both
al changes as well as global changes in the model. Examples of global changes are
scaling or shearing of the entire part in a single operation. A less obvious, but equally
portant aspect of editability is the system's ability to allow the user to make such changes
th global and local) interactively.

Brevity

gardless of how inexpensive available memory has become, the requirements of users
ve grown to make the compactness of the representation a critical issue. We intend
brevity attribute to reflect the size of the model, the amenability of the representation
various compression schemes and the ease with which modelling operations can be
formed on the compressed representation.

Analysability

alysability denotes the amenability of the representation to the following tasks in the
sign cycle: Computation/extraction of physical parameters from the model to validate
ainst specifications, functional analysis of the model (using finite element methods for
ample), and the visualization of the results of the analysis.

Reconstructibility

constructibility is the attribute that qualifies the ease with which volume representation
n be obtained from a physical part, using any one of several measurement modalities
g. range sensors, CAT scans, MRI scans etc.), or from a different volume representation
the same part.

0 Physical realisability

e usefulness of any volume representation scheme is considerably enhanced if a volume
del can be used to drive a fabrication process. In addition, the model may be used
determine various attributes of alternate fabrication processes. An example of such an
ribute is the time to machine a part. Physical realisability is the attribute of a volume
presentation that denotes these capabilities.

Volume modelling: The state of the art

this paper we examine four dominant modelling schemes, namely the boundary-repre-

(a)
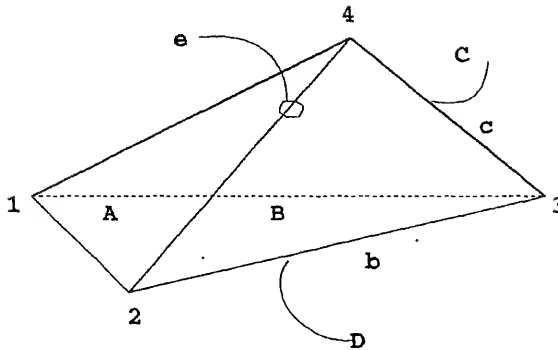
Edge	Vertices From	To	Faces Left	Right	Clockwise Pred	Succ	Counter Clockwise Pred	Succ
a	1	2	A	D	d	e	f	b
b	2	3	B	D	e	c	a	f
f	3	1	C	D	c	d	b	a
c	3	4	B	C	b	e	f	d
d	1	4	C	A	f	c	a	e
e	2	4	A	B	a	d	b	c

Figure 1. (a) Boundary-representation (B-rep) model.

The Boundary-Representation (B-Rep) scheme is a natural extension of the topological description of rigid solids, using complex data structures to represent vertices, edges and faces of a solid along with the topological constraints (such as adjacencies) and an elaborate system of pointers. A delicate balance has to be struck in B-Rep modellers to balance between efficiency of modelling operations and fast response to interrogations of the model usually implemented via redundant access paths. A thorny issue for B-Reps has remained the issue of robustness. Geometric computations and topological constraints in the framework of a redundant data structure can often lead to internal inconsistencies in the representation and to system crashes. Commercially available B-Rep systems
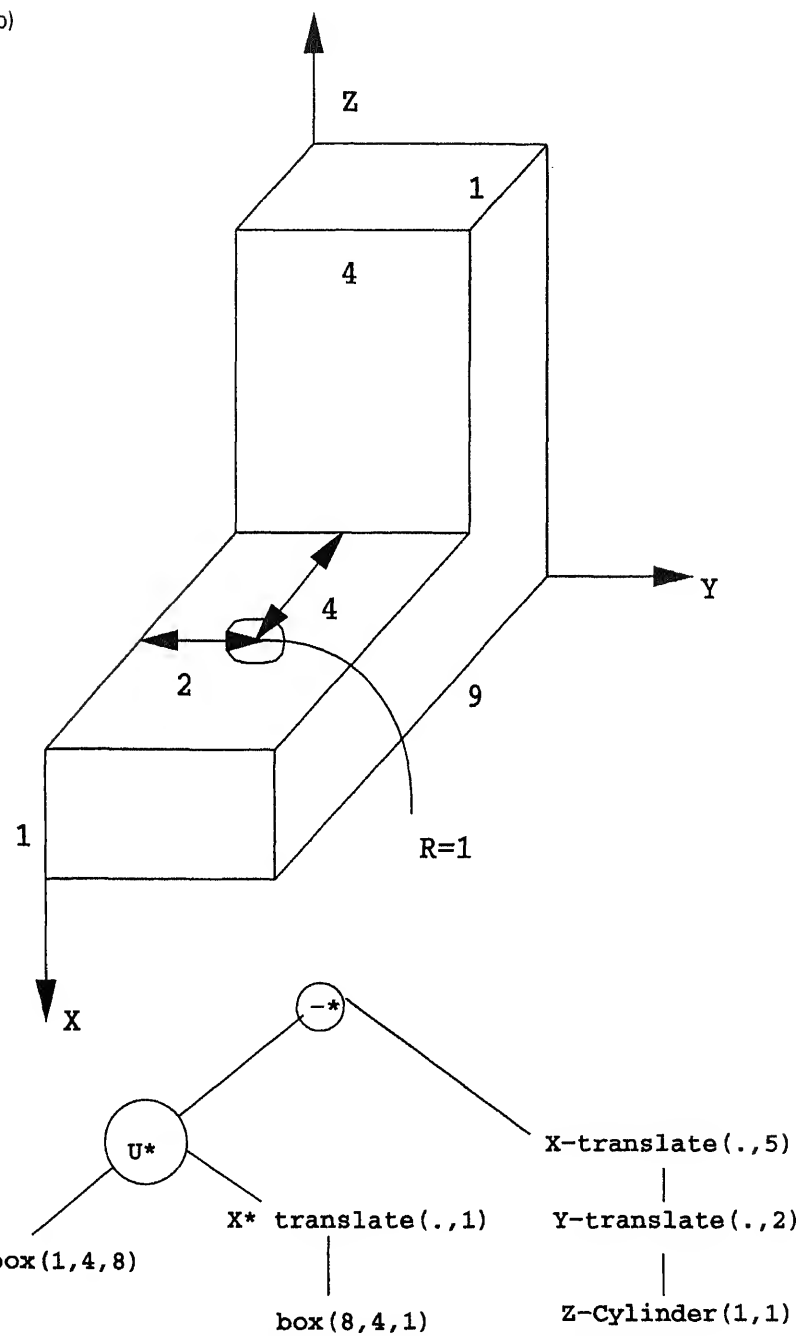
Figure 1. (b) Constructive solid geometry (CSG) model.
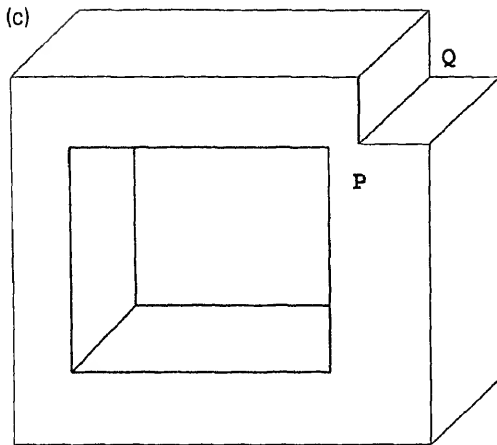
Figure 1.   (c) Non-manifold model.

popularized B-Rep modellers was the BUILD modeller, originally developed by I C Braid at Cambridge in 1979.

Constructive Solid Geometry (CSG) Modellers build representations of complex rigid objects by applying unary and binary operators such as set union, intersection, difference to elementary volumetric primitives such as blocks, cylinders, cones, spheres and tori. The representation is usually schematically shown as a tree with the root representing the solid and the leaves representing the primitives. The CSG modelling scheme was proposed and implemented (PADL-2) by a dedicated group of researchers at the University of Rochester (headed by H B Voelcker and A A G Requicha) in the early 1980's. The CSG representation is remarkable for its simplicity and succinctness and has therefore played an important role in the design of user interfaces for solid modellers. However, it is inadequate on some counts such as ease of rendering and analysis.

The Non-Manifold Representation was developed as an attempt to meet the need for a scheme in which the transition between 1D, 2D and 3D entities can be seamlessly executed. Dangling edges and faces, solids touching at vertices and edges and in the middle of faces, and other such irregular geometric constructs have to be valid in such a representation. Modellers built around such representations would be closed under general Boolean set operators and general sweep operators. Weiler (1988) proposed a radial edge data structure that could support such a representation. The approach that Weiler took was to extend the traditional B-Rep structure to admit non-manifold representations. SHAPES (XOX Corporation 1995) is a commercial non-manifold modeller developed and maintained by the XOX Corporation.

Voxel modellers, Spatial Occupancy Enumeration or Cellular Modellers use a very simple representation scheme. A volume of space (containing the object) is divided into a large number of cuboidal cells. The system then labels each cell as to whether or not it is occupied by material. The obvious data structure for this representation is a 3D array of 0's and 1's. More sophisticated data structures based on quadtree and octree representations have been implemented in the popular modellers TIPS-1 (built by Okino and others at Hokkaido University) and INSIGHT (a molecular modelling system built by Phoenix Data Systems). Pratt (1994) in his review of solid modelling, states that,"... most

d modelers based on a purely cellular approach are still in the research domain. This
ears to be because a satisfactory trade-off has yet to be found between the requirement
accurate boundary resolution needed in engineering applications and the very large
age requirements that this implies."

Requicha (1980) published a seminal study of solid modelling representation schemes.
more recent update of this work is by Pratt (1994). The interested reader is directed
hese papers for more details on the individual schemes. The above studies focus on
comparison of the B-Rep and CSG solid modelling schemes, while dismissing the
el-based modelling schemes as being impractical.

n this paper we re-evaluate these three schemes with reference to the desiderata listed
ow in table 1. As mentioned above, the non-manifold modelling paradigm evolved as
extension of the B-Rep representation scheme. However, in principle, non-manifold
dellers need not follow only this line of development. Hence we evaluate non-manifold
resentation schemes as separate from B-Rep of solids as also indicated in table 1.

A representation that flags geometric features of a part that helps identify manufacturing
cesses that match with a particular feature will enjoy tremendous increases in efficiency
he CAD system built around it. Such hierarchical representation schemes have come
be called *feature-based*. The need for them grew out of the fact that B-Rep and CSG
resentations are too low-level for a designer to work with. Feature-based schemes are
s an expedient interface between CAD systems today and analysis or manufacturing
lications (Rossignac 1990). We do not consider this scheme in our comparative study
volume representations since feature-based models exist at a higher level and can, in
nciple, be equally well-integrated with any of the B-Rep, CSG, Voxel or Non-Manifold
ume representation schemes.

Table 1 summarizes our evaluation of the various schemes for volume modelling via a
ng of each scheme on each of the ten attributes in our list of desiderata. The following
brief remarks on the ratings given in the table. The rating of voxel models will be
cussed in some detail in the sections that follow.

The boundary representation scores high on renderability, morphological dexterity, ac-
acy, editability (local and global), analysability, reconstructibility and physical realis-
lity. The detailed nature of this representation, with explicit access of vertex, edge and
e components, permits efficient data structures and computations on B-Reps. The ability

le 1. Attributes of the dominant modelling schemes.

ribute	B-Rep	CSG	Non-manifold	Voxel
nderability	Easy	Hard	Easy	Hard
rph. dexterity	Good	Poor	Good	Excellent
erogeneity	Impossible	Hard	Hard	Easy
curacy	Excellent	Poor	Excellent	Poor
tability (local)	Good	Poor	Good	Excellent
tability (global)	Good	Poor	Excellent	Poor
vity	Fair	Excellent	Fair	Poor
alysability	Good	Poor	Good	Good
constructibility	Good	Poor	Good	Excellent
ysical realizability	Easy	Hard	Hard	Easy

in principle to accommodate free-form bounding surfaces explains the high accuracy and morphological dexterity of this representation. On the other hand, the B-Rep scheme is verbose and has no provision to access volume elements in the interior of objects, and hence this representation scores low on the counts of brevity and heterogeneity.

Constructive solid geometry representations fare somewhat poorly in our evaluation. The only count on which we favour this representation is its brevity. This representation played an important role in the early development of solid modelling because it permitted approximate representation of complex shapes via regularized Boolean operations on simple primitives. However, the indirect and implicit form of this representation makes it cumbersome for interrogations that involve structure of the model.

Non-manifold models have even more detail explicitly given in their representation than B-Rep models. Hence they enjoy all the advantages of B-Rep models except on the count of physical realisability. The non-manifold representation allows features such as dangling edges and faces which are hard to fabricate. We rate non-manifold representations a little higher than B-Reps on global editability because of access to volume elements. On the negative side, non-manifold models suffer from low ratings on brevity and heterogeneity just as B-Reps do. In principle, because of access to volume elements, it is possible to work some heterogeneity into non-manifold models but this seems to involve considerable overheads.

3. Voxel modelling in emerging manufacturing technologies

In this section we focus on the attributes in which the Voxel methods outperform the other modelling schemes as shown in table 1. We believe that it is this set of attributes which can be exploited to tackle the challenging problems encountered by newer manufacturing technologies.

3.1 *Voxel modelling for virtual prototyping*

"Virtual prototyping is the process of design analysis, simulation and testing of a product within the computer and use of the results to refine the product before making a physical prototype" (Kumar *et al* 1996).

The excellent morphological dexterity of voxel models combined with their excellent local editability imply that CAD tools built using voxel models will allow arbitrary shape manipulation and sculpting. Such expectations have been fulfilled by interactive voxel modelling tools currently under development (Ravi 1996; Sethia 1996).

In VoxLab (Ravi 1996), it is possible to edit objects at the voxel level as well as use global operations to modify the volume structure of the object.

Interactive sculpting is the process by which a designer can impose arbitrary free-form shape changes on the object being designed. Considerable work has been done in the field of surface-based sculpting, which can be broadly classified into space deformation and surface fitting. Terzopoulos & Qin (1994, 1995) have developed dynamic NURBS and dynamic triangular B-Spline models which are physics-based models. These models are used for interactively sculpting curves and surfaces by applying simulated forces and local and global shape constraints. Pasko & Savchenko (1995) have approached the problem of interactive sculpting as the local deformation of a constructive solid by a set of arbitrary points

belong to the resultant surface. Rappoport *et al* (1995) have developed a method for modelling an object which is composed of several tensor product solids, while preserving desired volume of each primitive. Borrel & Rappoport (1994) have proposed a model producing controlled spatial deformations. In their model, the user defines a set of control points, giving a desired displacement, and radius of influence of each. Each constraint point, determines a local B-Spline basis function centred at the control point, and falling zero for points beyond the radius. Rappoport *et al* (1994) have introduced the concept soft constraints which need not be met exactly, thus increasing the design space. Finally, McCracken & Joy (1996) have proposed a new free form deformation technique that generalizes previous methods by allowing 3D deformation lattices of arbitrary topology.

The main drawbacks of these methods are that they are object-complexity sensitive, they do not operate on the object being sculpted directly, and it is very hard to modify the topology of the object being designed.

Voxel-based sculpting is proving its promise in our on-going work. We use Minkowski operations to sculpt voxel blocks using tools of arbitrary shape which are themselves voxel models (Sethia 1996). An object created by such sculpting operations is shown in figure 2.
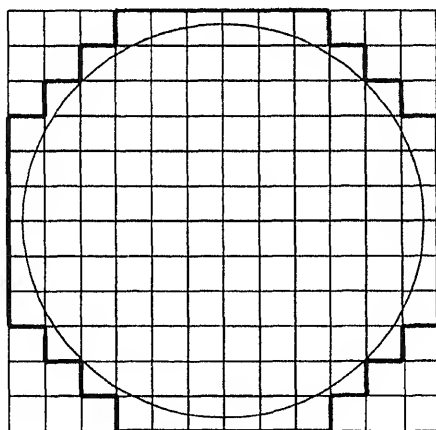
Figure 2. Spatial enumeration model

Mechanical analysis is an integral part of the design process since the physic properties of the parts determine if a particular design is viable or not. Traditional CA tools for analysis use the finite element method (FEM) to compute various material prope ties like structural strength, thermal conductivity etc. on the geometric representation of t part. The FEM methodology is essentially a technique for computing discrete approxim tions of continuous physical quantities. For B-Rep or CSG tree models the discretisati required by FEM is realized by tetrahedralisation of the design volume. A major co putational issue in this context is the maintenance of non-degenerate tetrahedralisati (ensuring connectivity, eliminating hanging edges etc.). These problems do not arise i voxel model since the basic unit of a voxel model is a volume-filling solid, namely, t voxel. Thus, accurate FEM discretisations can be easily extracted from a voxel-model f analysis.

The visualisation of the results of the analysis is facilitated by volume models: dir volume rendering techniques can be used to render the results of the analysis volumetrica superimposed on the voxel-model. Interactive volume visualisation tools allow the user view the results along any cutting plane of the model, or as a composite whole.

Several problems that are difficult using B-Rep or CSG or non-manifold approach have a simple solution using voxel models. These include estimation of mass properti interference detection, tolerancing and implementation of CSG operations.

Tolerancing is an important aspect of mechanical design (see ElMaraghy *et al* 1994, example). A voxel-based modeller can handle this aspect of the design as follows. T intersection of two objects due to the variation in tolerance is visualized as a 3D interfere volume itself. The interference volume is a measure of the extent of interference, a therefore a measure of the "out-of-tolerance" or deviation of the object. The toleran values can be reassigned, and a decreasing volume of intersection indicates a correct tre of tolerance alteration.

The use of interference volumes can be extended to solve the problem of assemblabil and disassemblability testing. Other modelling schemes require complex geometric int section algorithms and may also suffer from numerical inaccuracies. With voxel mode the interference can be precisely quantified and accurately located.

Combining the good analysability of voxel models with global and local editability is possible to design objects with excellent shape optimization. A naive approach is following: Map the results of the engineering analysis to the voxel model. Use a glo filter that removes voxels that do not have any load on them, after ensuring that th voxels are not essential for maintaining the structural integrity of the part. An exam of a shape-optimizing structure is a honeycomb mesh that can replace solid interior locations on the part that are not subjected to any loads. This will reduce the overall wei of the object as well as result in material saving. The morphological dexterity of vo models permits more sophisticated optimizations that will be impossible on B-Rep or C models.

3.2 *Voxel modelling for physical prototyping*

In spite of the tremendous strides made in virtual prototyping technologies, physi prototypes are indispensable in the concurrent engineering process. Hence attention

ng directed at making these physical prototypes as rapidly as possible. Traditional
totyping processes range from wooden mockups, casting from custom patterns, ma-
ning, and the use of sophisticated NC machines that are directly programmed from
CAD model of the part. *Rapid prototyping*, the ability to produce a prototype very
y in the design cycle, even before the tooling for the manufacture of the component
been designed, has emerged as a critical technological component of concurrent engi-
ring. Recently, several new manufacturing technologies that support rapid prototyping
e become available both in research laboratories and in the commercial marketplace.
se technologies are variously called *layered manufacturing*, *additive manufacturing*,
 stereolithography (Jacobs 1992). These manufacturing techniques allow a part, or a
totype, or a tool to be built by the gradual addition of material in a controlled way. In
trast, traditional manufacturing methods depend on the the removal of material (milling,
ning etc.) or on the deformation of material (casting, moulding etc). The field of rapid
totyping is making rapid strides (Burns 1993).

RPTs provide the means to manufacture parts that cannot be realized by conventional
chods (Kruth 1992). Arbitrary shapes, composite materials, and complicated geometries
 be as easily manufactured as simpler geometries. In order to exploit these capabilities,
modelling schemes used by CAD tools to design parts should be capable of handling
nge of shapes including non-manifold objects. Thus, non-manifold schemes (XOX
poration 1995) and voxel-modelling are the only schemes that have this capability.

However, for converting a virtual prototype into a physical prototype using RPT tech-
ogy, the current practice is to convert all models into the stereolithography format (STL
nat) that has emerged as the *de facto* standard. The model in STL format, basically
D triangulation, is then input to a slicing process (that is unique to each of the lay-
d manufacturing equipment) that generates a sequence of slices that drive the material
ition process.

These two steps, namely, converting into a STL format and the slicing of the STL format
oduce changes in the model that are beyond the control of the part designer.

Using a voxel-based modelling scheme it is possible to completely eliminate the above
 steps and instead drive the fabrication process directly. This is facilitated by the close
respondence between voxel models and the fabrication processes.

The current range of rapid prototyping machines can be broadly classified into two major
egories based on the way in which material is added to an object under construction.

Parallel/image-based: The most prominent system under this class is the Solider MPM
ystem of Cubital, which is a commercially available system. A well-known research
ystem under this category is the MD* (Weiss & Prinz 1991). In these systems, successive
ayers of the component under fabrication are created by the use of masks that either
llow a light source to solidify a photopolymer under the exposed regions, or spray
naterial on the exposed parts of the mask. The advantage of this approach is that the
ime for creating a complete layer does not depend on its geometric complexity. Each
nask is simply a slice of the object, and can be thought of as the image of the object's
ross-section.

this category. In these systems, a layer is formed by the sequential formation (deposition, or solidification) of the contours in the object's cross-section. Hatching or filling-in operation is needed to obtain solid interiors.

Some RPTs, still in the research stage, like the ballistic particle manufacturing, and the beam interference solidification process, create parts by solidifying individual voxels.

Even some layered manufacturing techniques like stereolithography and selective laser sintering use point-by-point solidification to grow each layer. Thus these processes work one level below the layer-by-layer manufacturing techniques and this involves the conversion of layer information to voxel information, usually performed by the manufacturing equipment itself.

At every stage in the design process, the description of the part is directly the description required by the manufacturing process. Thus it is easy to evaluate the design with respect to its manufacturability.

Several current and proposed RPTs generate parts by adding material one layer at a time. The MD* process (Weiss & Prinz 1991) uses this approach using photographic masks and a vapour deposition process that is very similar to the technology used for VLSI fabrication. However, the input to the manufacturing device is currently obtained by a postprocessing step after the design of the part is completed. This step involves the generation of slices of the finished part, the number of slices generated depending on the layer resolution of the RPT. The hardware itself is neutral to the CAD tools or primitives that have been used in the design of the part. However, the use of traditional CAD tools that use either B-Rep or CSG as the basis will imply that the postprocessing step is left to deal with problems like the proper orientation of the part before slicing for optimal manufacturability (minimum use of support material, protection of the structural integrity of the part at every stage of the growth process etc.).

STL (Stereolithography list) was developed by 3D System for their stereolithography machine and has become a *de-facto* standard input format for several RPTs (Jacobs 1992). This PHIGS-based format approximates the part's inner and outer surfaces by a set of triangular plane patches. Each patch is described by its three vertices and a normal vector pointing out of the material.

Though it is possible to generate STL from voxel models to conform to existing RPTs, the voxel models permit generation of outputs in formats more closely matched to the target RPT. For instance, if the RPT uses a layer-by-layer approach, the voxel-based modeller can output the layer information after taking into account several process-planning steps, even during the design phase. This will eliminate the need for any processing by the RPT.

Composites: Research in LM technologies will make it feasible to fabricate components with several constituent materials that are physically located to ensure optimum performance. The extrusion process used by the Fused deposition modelling (FDM) of Stratasys (Burns 1993) can use polymers as well as metals of low melting point like tin. The MD* process of CMU uses a robotic arm to spray different metals over masks. It is very likely that in the near future, LM technology will mature to a point where multiple materials can be used to fabricate a single component.

The ease with which heterogeneity can be handled by voxel models enables designers to

Conventional design tools are not oriented to the design of composite objects. Specialized tools are used in areas like the aircraft industry where composite materials play a major role.

A voxel-based modeller can ultimately provide the capability to design a composite object with materials selectively placed at individual voxels. There is no need to compute the complex geometries of the interleaved materials since, for fabrication, each slice of the voxel buffer can be directly read out and several masks per layer can be created to deposit the different materials. Such capabilities will be indispensable as the technology of microelectromechanical systems (see Senturia *et al* 1992, for example) matures: A voxel will then be of molecular dimensions.

Inspection of the fabricated part to ensure its compliance with specifications is an important aspect of manufacturing. A voxel-based approach has been successfully used to implement an on-line automated visual inspection system (Tarbox & Gottschlich 1995).

Fabrication of components with embedded electronics will be an increasingly common requirement. The design of wearable computers is one such example that requires rapid product development (due to the rate of change of the component technologies) and prototyping. A simple approach that exploits volume models in combination with layered manufacturing technology is as follows: Volume-scan the electronic subunit, combine with a voxel model to give an integrated model of the component and insert the electronic unit at a suitable time during fabrication. Implementation of this basic process requires advances in both CAD tools and layered manufacturing tools.

Chandru *et al* (1995) elaborate on the above advantages of using voxel models for rapid prototyping.

3.3   *Voxel modelling for reverse engineering*

Voxel models have excellent reconstructibility. This implies the following.

- It is easy to obtain the voxel model from a physical part. In practice, CAT and MRI scans directly give a three-dimensional voxel array of the model while range sensors or 3D scanners provide a cloud of points from which a voxel model can be obtained.

- It is easy to convert models from other representations to voxel models. This can be achieved using a class of algorithms called voxelisation algorithms (Wang & Kaufman 1993) that convert geometric primitives into an array of voxels.

Reverse engineering is the process of construction of a CAD model from a physical part. Measurements are made on the physical prototype (typically by scanning) and these are then processed to obtain the CAD model. Thus voxel models are ideally suited to support reverse engineering.

In addition to the traditional reverse engineering task of obtaining a computer model from a physical prototype, the voxel-based approach leads to other possibilities. Reverse engineering in conjunction with the voxel-based modelling and analysis provides a means of reengineering a part from a sample physical model.

We are investigating the use of such reverse engineering in conjunction with rapid prototyping equipment to provide a platform for rapid product innovation and design

modelling and volume-error measures.

4. Research issues in voxel modelling

Table 1 lists some attributes for which voxel models compare poorly with the other models. These are precisely the attributes that have been the bottlenecks to the widespread use of voxel models. In this section we address each of these attributes and indicate research directions aimed at overcoming the limitations.

Rendering: Current graphics systems are predominantly polygon-based. Recent desktop workstations, even at the entry level, are capable of rendering close to a million triangles per second. Similar hardware advances are yet to take place for volume rendering and hence rendering of reasonable voxel volumes (say $512 \times 512 \times 512$) takes a few seconds. However, various ways to speed up volume rendering are being vigorously pursued. These include the use of texture mapping hardware to perform back-to-front composition of volume slices (Fraser 1994), the development of special-purpose hardware for volume rendering (Pfister *et al* 1995), and the use of distributed and shared memory workstations.

Global editing: Voxel modelling has excellent local editability and certain global operations (analogous to filtering) that alter the entire model can be implemented with ease. However, they suffer from a lack of information about the component structures of an object. For instance, a voxel model of a table will allow one to edit the sharp corners into rounded corners, but it will be difficult to replace the four legs of the table with legs of different shape. This is because there is no notion of a substructure in the basic voxel models.

Research into hierarchical models that permit such substructuring, without disturbing the other advantages of voxel models may be thought of as feature-based modelling using the voxel as the basis. This appears to be a fruitful research direction.

Brevity: A serious practical problem with voxel models is their poor brevity. A regular voxel array to represent a component with a 5 cm bounding box will require a voxel resolution of about 400^3 to realize a physical part with an accuracy of 125 microns.

Clearly, this is another reason for hierarchical and adaptive representations. Octrees, BSP trees and similar data structures combined with algorithms that effectively operate on these structures is a critical area of research. Another facet is that of compression algorithms and algorithms that operate on compressed data structures.

Accuracy: The accuracy of voxel models is inversely related to the brevity of the representation. Since by definition a voxel representation is a discretization of continuous geometric entities, the accuracy of the representation will be poorer when compared to CSG or B-Rep. Aliasing errors due to the discretization process will be present. However, the redeeming aspect of voxel models is the possibility of exact quantification of the error as well as user control of the error. Research into the application of anti-aliasing techniques of computer graphics to voxel models and the use of volumetric error measures is a promising area of research (Prakash & Manohar 1995).

rences

l P, Rappoport A 1994 Simple constrained deformations for geometric modelling and inter-
ive design. *ACM Trans. Graphics* 18: 137–155

M 1993 *Automated fabrication* (Englewood Cliffs, NJ: Prentice Hall)

lru V, Gurumoorthy B, Manohar S 1996 Integrated modelling and reverse engineering
layered manufacturing. Technical Report, CSA-IISc-1996-02, Indian Institute of Science,
ngalore

lru V, Manohar S, Prakash C E 1995 Voxel-based modelling for layered manufacturing.
EE Comput. Graphics Appl. 42–47

raghy W H, Valluri S R, Skubnik S M, Surry P D 1994 Intersection volumes and surface areas
cylinders for geometrical modelling and tolerancing. *Computer-Aided Design* 26: 29–45

R 1994 Interactive volume rendering using advanced graphics architectures. *SGI Dev. News*
cember

r S *et al* 1996 Rapid design and manufacture of wearable computers. *Commun. ACM* 39(2):
–70

ann C M 1989 *Geometric and solid modelling: An introduction* (San Mateo, CA: Morgan-
ufmann)

s P F 1992 *Rapid prototyping and manufacturing – Fundamentals of stereolithography.*
earborn, MI: Society of Manufacturing Engineers)

J P 1992 New manufacturing techniques for rapid prototyping and concurrent engineering.
nufacturing in the era of concurrent engineering (eds) G Halevi, R D Weiel (New York:
sevier Science)

r V, Bajcsy R, Harwin N, Harker P 1996 Rapid design and prototyping of customized
abilitation aids. *Commun. ACM* 39(2): 55–62

acken R, Joy K I 1996 Free-form deformations with lattices of arbitrary topology. *Computer
aphics Proceedings, SIGGRAPH96*, pp 51–60

ara J *et al* 1992 Virtual space decision support systems and its application to consumer
owrooms. Matsushita whitepaper

A A, Savchenko V V 1995 Algebraic sims for deformation of constructive solids. *Third
mposium on solid modelling and applications*, pp 403–408

r H, Kaufman A, Wessels T 1995 A scalable architecture for real-time volume rendering.
oceedings of the Eurographics Workshop on Hardware*, Masstricht, pp 123–130

sh C E, Manohar S 1995 Error measures and 3D Anti-aliasing for voxel data. *Proceedings
Pacific Graphics 95* (Seoul: World Scientific)

M J 1994 Solid modelling – Survey and current research issues

oport A, Hel-Or Y, Werman M 1994 Interactive design of smooth objects with probabilistic
int constraints. *ACM Trans. Graphics* 18(2): 156–176

oport A, Sheffer A, Bercovier M 1995 Volume preserving free form solids. *Third symposium
solid modelling and applications*, pp 361–372

N 1996 *VoxLab: An interactive toolkit for voxel-based modelling*. Masters Project Report,
partment of Computer Science and Automation, Indian Institute of Science, Bangalore

icha A A G 1980 Representation of rigid solids – Theory, methods and systems. *Comput.
rv.* 12: 437–464

icha A A G 1984 Representation of tolerances in solid modelling: issues and alternative
proaches. *Solid modelling by computers* (New York: Plenum)

gnac J R 1990 Issues on feature-based editing and interrogation of solid models. *Comput.
aphics* 14: 149–172

Senturia S D *et al* 1992 A computer-aided design system for microelectromechanical systems (memcad). *J. Microelectromech. Syst.* 1(1): 3–13

Sethia S U 1996 *Interactive volume sculpting*. Masters project report, Department of Computer Science, Indian Institute of Science, Bangalore

Terzopoulos D, Qin H 1994 Dynamic NURBS with geometric constraints for interactive sculpting. *ACM Trans. Graphics* 18(2): 103–136

Terzopoulos D, Qin H 1995 Dynamic manipulation of triangular B-splines. *Third symposium on solid modelling and applications*, pp 281–291

Tarbox G H, Gottschlich S N 1995 IVIS: An integrated volumetric inspection system. *Comput. Vision Image Understanding* 61: 430–444

Wang S W, Kaufman A 1993 Volume sampled voxelisation of geometric primitives. *Proc. of IEEE Visualization 93* (New York: IEEE Press)

Weiler K J 1988 The radial edge structure: A topological representation for non-manifold geometric boundary modelling. *Geometric modelling for CAD applications* (ed.) M J Wozny *et al* (Amsterdam: North-Holland)

Weiss L E, Prinz F 1991 A framework for thermal spray shape decomposition in the MD* system. *Proc. of 1991 Solid Freeform Fabrication Symposium*

XOX Corporation 1995 SHAPES microtopology, a system for modelling mesh geometries. XOX Corporation, 1450 Energy Park Drive, Suite 120, St. Paul, MN 55108

ature-based geometric reasoning for process planning

G ADITYA NARAYAN, S R P RAO NALLURI and B GURUMOORTHY

Department of Mechanical Engineering, Indian Institute of Science, Bangalore 560 012, India
e-mail: bgm@mecheng.iisc.ernet.in

Abstract. We present a framework based on Domain Independent Form (DIF) features for automatic evaluation of manufacturability and process planning for machining. The framework enables interpretation of a common product model with respect to each task in the transition from design to manufacture. A key idea here is to generate the interpretation suitable for each task in two steps. In the first step, DIF features that are defined through feature enumeration are automatically extracted from the geometric model. The extracted DIF features are then mapped into features meaningful for individual tasks through geometric reasoning based on domain dependent knowledge. The formal approach to feature definitions and separation of the domain specific reasoning from the general geometric reasoning enable us to overcome the bottlenecks reported in features technology.

Keywords. Feature-based manufacturing; feature extraction; feature mapping; visibility; process planning; manufacturability evaluation.

Introduction

e concept of feature evolved from the desire to devise methods for integration of part metric model with applications such as process planning, group technology coding and nerically controlled (NC) tool path programming. Features are high level abstractions part geometry such as holes, slots and ribs around which engineering knowledge and ertise is structured. Hence reasoning about part geometry in applications such as process nning becomes easier to automate when part description becomes available in terms of tures rather than basic topological entities (face, edge and vertex) that are available in geometric model. These features are referred to as *form features* as they describe the m and shape of a part.

The motivation for integrating part geometric model (also referred to as the CAD model) h applications such as planning for manufacture, comes from the need to compress prod- development lead times and to reduce costs. The sequential nature of product design l the necessity of human intervention in the transition between individual tasks have n responsible for large lead times and costs in the conventional product development

217

cycle (Bakerjian 1992). By allowing downstream lifecycle concerns to influence product design during the design process itself, delays and costs due to the sequential nature of the development process can be reduced considerably, if not eliminated. In the context of products manufactured by machining, this would imply ensuring that the design is machinable during the design process itself. Eliminating human intervention in the transition between tasks involves automating the geometric reasoning involved in interpreting part geometry appropriate to the task domain and the process-based reasoning to realise the task.

Automation of evaluation of manufacturability and planning of manufacture requires that part information be available at different levels of abstraction and detail. For instance the task of generating a process plan involves sequencing of machining instructions (gross process plan) and generation of NC tool path (fine process plan). The first requires information on entities such as slot, step or hole, whilst the second also needs the geometric details such as surface normals and locations. The domain knowledge/expertise of each task in product development is structured around features. Hence it is essential to support features to automate the reasoning involved in evaluation of a design with respect to a task.

There have been three routes followed by researchers to realize feature information to reason about process (Shah 1991). These are as below.

• Interactive feature definition: In this approach, features are defined by the users by picking entities, associated with each feature, from a CAD model. This approach has been used mainly to input data to process planning and tool path generation programs (Chang & Wysk 1985).

• Feature-based modelling (FBM): In this approach the feature information is incorporated in the model construction stage itself. This is achieved by using features to build the model (Salomons *et al* 1993). This method of modelling is a generalisation of the 'design by features' approach. In the 'design by features' approach only design features are used, whereas in FBM, features of interest to any one specific application can be used. In most implementations, feature-based modelling is based on Constructive Solid Geometry (CSG), where the features are available as primitives in addition to the regular primitives such as cylinder and block. Commands of a solid modeller to create the feature are generated based on the feature selected. Arbab (1992) modelled parts by removing volumes corresponding to machining features from the stock solid. Part feature model thus obtained was used to link design with process planning. Shah (1988) implemented a general purpose feature-based modelling system. This system allows the user to define a set of features along with all its manipulation functions.

• Automatic feature extraction: In this approach, form features of interest to an application are automatically extracted from the geometric model of a part. Automatic feature extraction has fundamental significance from the integration perspective of product development cycle. Reasoning involved in the downstream applications cannot be automated without correct interpretation of part design with respect to each task. Since most of the established application knoweldge is based on human interpretation of part geometry, it is essential to generate human-like interpretation for complete automation.

Automatic extraction of features from a geometric model is a difficult problem. Application and context dependence of features make it doubly difficult. The basic process

have been proposed to solve this problem. These vary from syntactic pattern recognition techniques (Kyprianou 1980), graph matching (Joshi & Chang 1988), and volume-decomposition (Shah *et al* 1994; Tseng & Joshi 1994; Sakurai & Dave 1996) to hint-based techniques (Vandenbrande & Requicha 1993). The reader is referred to Shah (1991) and Subrahmanyam & Wozny (1995) for reviews of the literature on feature recognition.

None of the commercial CAD/CAM/CAE software that claim to be feature-based can integrate design with manufacturing, analysis with other downstream tasks. Most of these software support only the FBM approach. In this approach a task-specific model can be obtained only by restricting the user to models using a particular set of features. For other tasks, the part has to be modelled differently and hence both data integrity and development time are adversely affected.

On the other hand, feature recognition approach to generate feature models is only now beginning to see commercial development (CAD 1997). Most of the techniques reported in literature have limited scope and their extension to handle real world part complexity is not proven. Implementations reported so far consider only one application (process planning for machining) and domain specific heuristics are used to recognise interacting features (Subrahmanyam & Wozny 1995). The most formal and general approach is graph matching, but sub-graph isomorphism is in NP-complete list. This approach also suffers from combinatorial explosion problem as sub-graphs of all feature instances have to be defined in advance and checked for. Techniques based on decomposition (Shah *et al* 1994; Tseng & Joshi 1994; Sakurai & Dave 1996) are proving to be more successful than graph-based approaches in handling feature interactions, at least for simple cases.

In the proposed framework, *domain-independent form* (DIF) features are used as the intermediate representation. The term 'domain-independent' is used to imply that the definitions and classification of the features are based purely on form (geometry and topology) and not on considerations arising from various application domains such as machining, near-net shape processing, analysis, or assembly. The DIF feature set is then used to obtain a feature set corresponding to other application domains through a process called feature mapping. In the following we first present the feature-based framework for manufacturing (in particular, machining). Definitions for DIF features are presented next. We then highlight some of the results in feature extraction, mapping and manufacturability evaluation. We conclude with some open issues in this area.

2. Feature-based manufacturing framework

A product typically consists of several electrical, electronic and mechanical parts. The framework proposed addresses only the development of discrete mechanical parts. Part development involves more than one task, and even in a single task part data may have to be interpreted in more than one way. For example, if a part has to be manufactured through machining, then it has to be considered for process planning, which in turn consists of several sub-tasks such as stock selection, fixturing, sequencing, tool selection, NC tool path generation and final inspection.

Generating part interpretations suitable to each task through FBM approach would involve modelling the same part differently for each task. This would adversely affect the
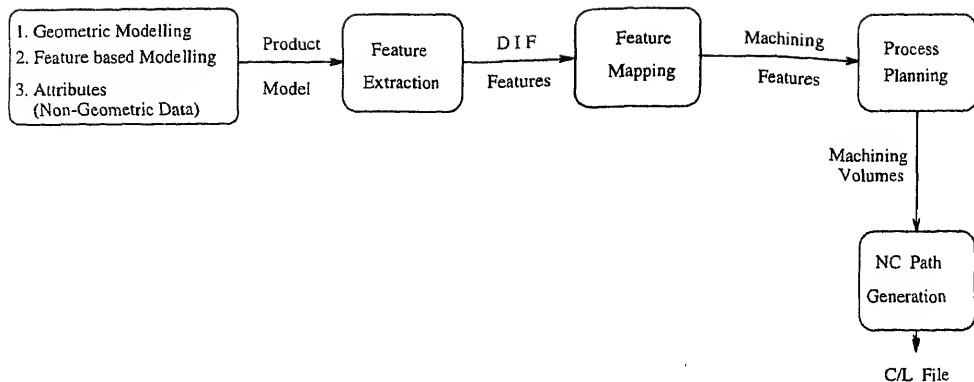
Figure 1. Feature-based machining system.

concurrency and development time. On the other hand, implementing feature extraction specific to each application involves bringing down the level of abstraction of each application to the level of basic geometry, which is a difficult task. Moreover, automatic handling of non-geometric data which can only be attributed to features is not clear at present in the feature recognition approach.

The proposed framework is designed such that feature-based modelling (FBM) and feature extraction approaches complement each other (see figure 1). Migration from geometric modelling to product modelling is effected through providing features for modelling. Product data which are difficult to attribute to basic geometric entities can be attributed to features. Unlike the conventional FBM approach, the designer is not restricted to use only features for modelling and can use boolean operations over primitives along with features. As features get modified in the process of modelling, the design model does not correspond to the feature model of the final solid. In the proposed approach, part geometry is represented as boundary representation (B-rep). Product data are represented as attributes to entities in the design model and geometric model.

Design model is an unevaluated representation of the design process. Reasoning over design model to generate task-specific interpretations is difficult because of its subjectivity (a designer's viewpoint) and non-uniqueness. Hence the design model is referred only for product data in the mapping process. From the evaluated geometric model (B-rep), DIF feature model is created through automatic feature extraction. Then the extracted DIF features are mapped into each specific task. DIF feature interpretation is at a higher level of abstraction than the basic geometry. Hence, knowledge-based reasoning involved in mapping can be implemented without bringing down the domain knowledge of each task to the level of basic geometry.

We have taken a general and formal approach to DIF feature classification and definitions. Based on this formalisation, problems in feature recognition and feature-based modelling are treated generally without using any one domain specific heuristic. All the task specific issues are addressed in feature mapping and subsequent process specific modules, process planning and NC tool path generation (figure 1), where domain knowledge interacts with the part's DIF interpretation. Process-based reasoning has been divided into three modules, as the interaction of process and part geometry is of different degrees in
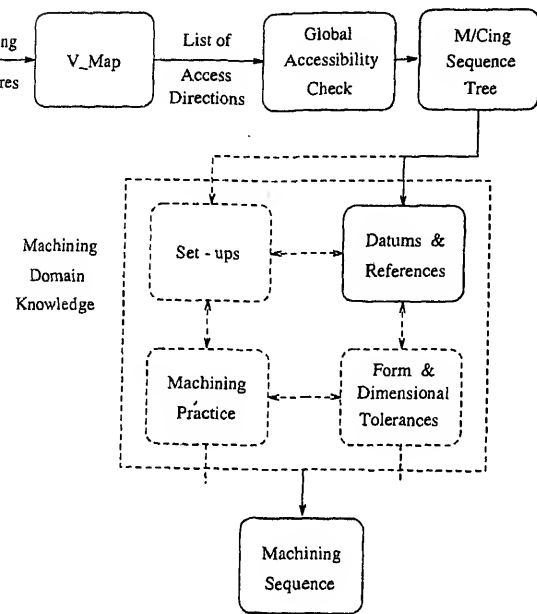
Figure 2. Manufacturability evaluation and planning for machining.

h of the modules and the nature of reasoning is also different. For example, in NC
l path generation the task primarily involves geometric computations, once the cutting
ameters are fixed in the planning module.

n this paper, we present a feature-based approach for the planning module where the
ective has been to use geometric reasoning to provide plans that can then be pruned by
istraints from the process or processing environment. This is illustrated in the schematic
figure 2. The output from the mapping module of figure 1, machining features, forms
input to this module. Module V_Map uses visibility checks on the faces of each feature
generate access directions for machining the feature. A global accessibility check along
feasible set of the access directions is then used to evaluate manufacturability of the part
l if the part is manufacturable, to generate all possible sequences for machining the part.
is *sequence tree* is then pruned by subjecting it to constraints from the machining domain
h as set-up optimisation, handling datums/references, existing machining practice to
tain one or more machining sequence. The remainder of the paper explains this approach
l provides some results.

Domain-independent form features

th feature extraction and feature-based modelling require a set of well-defined form
tures. A feature has to be defined either interactively or through a language before it
1 be used either by feature recognition or by feature-based modelling modules. Form
tures unique to certain applications such as twists, louvres, tabs of sheet metal, and
ondary features such as knurl, teeth and threads interact with the existing part geometry
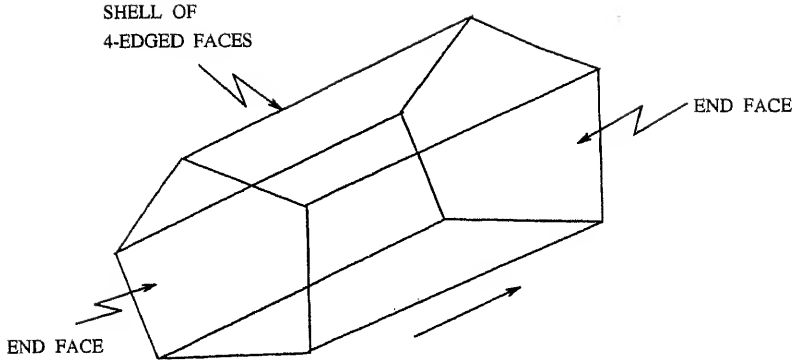
CEF	10
SEF	3
SSFs	1, 2
CSFs	7, 8, 9
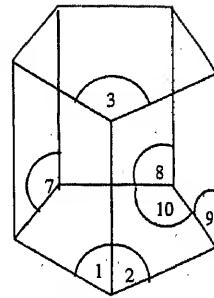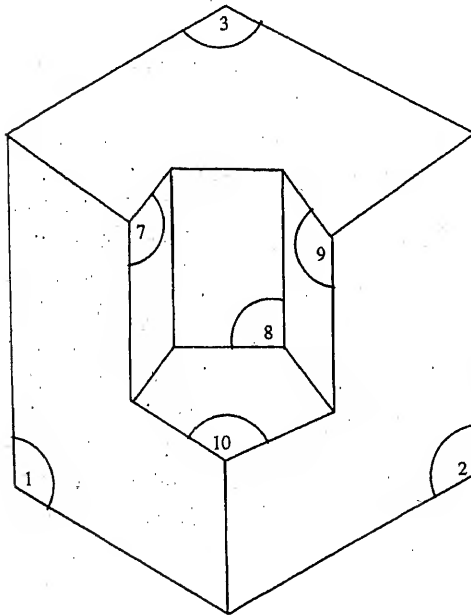
Figure 3. Sweep structure and face classification for DIF features.

depends on the application, context, and interactions. Hence we focus on classification, extraction, mapping and manipulation of primary (non-unique) features.

The conventional approach (Butterfield *et al* 1986) is first to collect form features of interest to a particular application. Collected features are then organised through classification into hierarchies and feature definitions are derived. In this approach, the number of form features collected and the criteria used for classification are subjective. Hence, finiteness and completeness of the defined feature set cannot be established with respect to possible variation of topology and geometry in designed components. Unlike earlier

le 1. Basic DIF feature types.

of l faces	No. of end faces			
	No ends (closed)	0 (Double blind)	1 (Blind)	2 (Through)
les)	Closed hole	Double blind hole	Blind hole	Through hole
ts)	Closed slot	Double blind slot	Blind slot	Through slot
djacent ner slots)	Closed corner slot	Double blind corner slot	Blind corner slot	Through corner slot
on-adjacent tual slots)		Double blind virtual slot	Blind virtual slot	Through virtual slot
ee or more than ljacent tual corner slots)		Double blind virtual corner slot	Blind virtual corner slot	

roaches we have defined features through enumeration of possible topological varia-
ns in designed solids.

We have modelled form feature-generation as subtraction/addition of a solid_piece
m/to a base_solid (Rao Nalluri 1994). The solid on which the feature is created is
erred as the base_solid and the minimum solid that is required to create the feature
referred as the solid_piece. In the generation of a feature, some existing faces of the
e_solid are modified and some new faces are created. The modified and the newly
ated faces are classified as 'shared' and 'created' faces respectively w.r.t. the generated
ture. The solid_piece faces can be classified into 'shell' and 'end' faces, and combining
two classifications results in four face types: Created Shell Face (CSF), Created End
ce (CEF), Shared Shell Face (SSF) and Shared End Face (SEF) as shown in figure 3.
e geometry and arrangement of 'created' faces of the feature is defined as the 'shape'
the feature. The connectivity of the feature face set with the rest of the base_solid is
ined as 'type' of the feature. In the feature generation model, keeping the variations
ich affect the 'shape' independent, variations which affect 'type' are enumerated and
eric feature definitions are derived in terms of the four face types as given in table 1.

Direct manipulation through form features

ature-based modelling (FBM) provides an intuitive design environment and enables at-
ution of product data to the features. The common approach to implement FBM system
o have features supported only at the user interface. For all manipulations of geomet-
representation, features are converted into solid primitives and boolean operations.
contrast to earlier approaches we have developed algorithms for direct manipulation
ao Nalluri & Gurumoorthy 1993a) of boundary representation (B-rep) through fea-
es. Feature generic 'type' dictates the manipulation involved in creation and deletion
a feature instance. All feature-based editing of B-rep are effected through deletion of
old feature and creation of a modified feature. Feature relations are captured through

shared/created face classification. Automatic adjustment of features in editing is effec
based on the captured feature relations. Manipulation of B-rep directly with featu
through computationally efficient incremental upgradation distinguishes our work fr
other efforts in this area.

5.   Extraction of DIF features

Extraction of DIF features involves decomposing of the input solid into a set of instan
of enumerated generic form features. We have modelled the feature extraction process a
reverse process of feature generation. In the process of feature generation the topologi
complexity of the base_solid increases. Hence in the process of feature extraction part s
should get simplified with extraction of each feature. The idea of topological simpli
is implemented through the concept of Dynamic Topological Status (DTS) (Rao Nal
1994; Hari 1995). DTS of a face is a set of topological factors about the face which descr
its topology. Since features are created in a certain order there is an order in their extracti
A feature gets extracted when all its faces are only 'created' faces and not the 'shar
faces of any other features. 'Created' faces have simple topology compared to 'shar
faces. DTS values of faces enable clustering of 'created' faces of a feature and also g
clues about the probable 'type' of the feature. The probable feature 'type' is confirm
based on the enumerated feature definitions. For example, the clue for a blind_slot feat
is a face with all 4-edged neighbours except one corresponding to single shared shell f
(SSF). The confirmation procedure for feature type blind_slot is as given below. Once
'type' is fixed, the feature face set is identified and fixing 'shape' becomes a local match
problem.

Procedure. Blind-slots(face[i])

    Feature-Type := NULL
    If (face[i] DTS1 != 4 and DTS2=1 and DTS3=0)
            End-face1:=face[i]
        `   Find the single non-four-edged neighbor of End-face1
            If (It is of type SSF)
                    Find End-face2
                    IF(DTS of End-face1 and End-face2 are IDENTICAL)
                            IF(Angle btw End-face1 and CSF's is CONVEX)
                                    Feature-Type:= Double Blind Slot
                            ELSE
                                    Feature-Type:= Double Blind Rib
                    ELSE
                            IF(Angle btw End-face1 and CSF's is CONVEX)
                                    Feature-Type:= Blind Slot
                            ELSE
                                    Feature-Type:= Blind Rib
    Return(Feature-Type)
End Procedure

The separation of 'type' and 'shape' in feature definitions reduced the number of confirmation procedures required from $(n_t \times n_s)$ (n_t = number of types, n_s = number of shapes) to $(n_t + n_s)$. The complexity of the search depends on the number of faces rather than the number of feature 'types' supported. Hence the complexity of feature extraction is $O(n_f^2 \log(n_f))$. Since all feature interactions are covered in enumeration for defining features, interactions do not require any special heuristic procedures for their resolution. The order obtained in the extraction can be used in mapping.

Inability to deal with interacting and intersecting features in a general way is one of the bottlenecks in features technology. We attribute this problem to (1) incompleteness of the feature set w.r.t. to topological and geometric variation in parts, (2) attempting to recognise a task specific feature when it is not explicit through relaxation of feature definition and application of domain-specific heuristics. The formal approach to feature definitions enabled us to define a complete set of feature types that is required to cover the defined topological and geometric variation. In the case of intersecting features, as we do not attempt to recognise the task-specific features directly, the most explicit feature is recognised without any application concern. Later in the mapping, based on the relationships among extracted DIF features, application specific interpretation is obtained.

6. Mapping DIF features to applications

Mapping involves knowledge-based reasoning along with geometric reasoning. Relations among extracted DIF features are explicitly represented as Feature Relational Graph (FRG) to apply the high level application knowledge. FRG is constructed based on the shared/created face classification. A feature X is said to be a child of another feature Y, if and only if, at least one shared face of feature X corresponds to a created face of feature Y. FRG is an acyclic graph which stores the child/parent relationships among features. Mapping modules have access to (1) geometric model, (2) FRG of extracted DIF features, and (3) design model. Both (1) and (2) representations are domain-independent, and entities in representation and their interpretation are well defined, whereas the design model is a designers interpretation and hence is subjective. It is mainly referred to get the task-specific product data. These three interpretations at different levels of abstraction provide all the information at the required levels to enable knowledge-based reasoning in mapping.

The inputs to the mapping module are stock B-rep, part B-rep, and FRG. The feature extraction module classifies the DIF features extracted based on shape (rectangular, circular, I-section etc.), class (hole, slot, step) and type (positive, negative). With reference to the feature definition, positive features and negative features are obtained by adding and removing material respectively from the base_solid. Negative features without positive children map directly into machining domain. Positive features that are child features of a negative feature are split into the portion totally inside the negative feature volume and the portion outside. The portion outside the negative feature volume is treated as a separate positive feature then onwards. The portion of the positive child feature inside the negative feature and the negative feature are combined as a single negative machining zone. Positive features which are adjacent are clustered and enclosed with a box. Positive feature cluster and box together form a positive machining zone. The boxes corresponding to positive
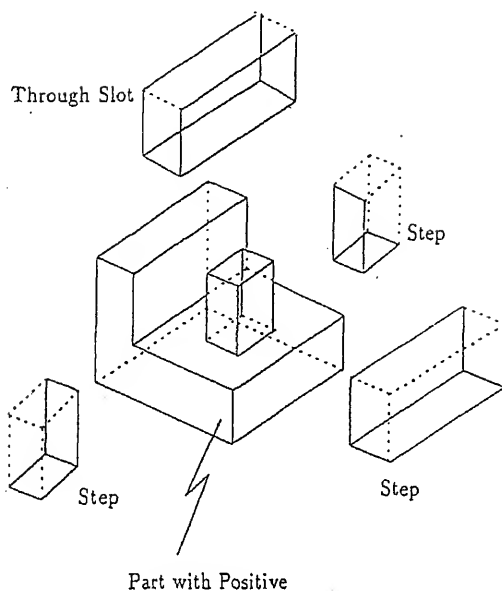
Figure 4. Mapping DIF features to machining features.

machining zones and base_solid make the first stage solid. First, stock solid is reduced first stage solid through clipping (Rao Nalluri & Gurumoorthy 1993b; Rao Nalluri 19 then enclosing box in each positive machining zone is clipped w.r.t. inside feature clu In clipping a outside solid w.r.t. an inside solid, the outside solid is split about the f of the inside solid. The results of clipping are convex volumes that are required to subtracted from the outside solid to get the inside solid. After the clipping, all posi features and the base_solid are created in the stock solid. As all negative features hav be either on base_solid or on one of the positive features, they can be machined any after their shared faces are ready. Through clipping all part positive features are conve into negative machining features w.r.t. stock. These negative features are added to the features list for further processing. Figure 4 shows a part with positive DIF feature an negative features obtained after mapping.

7.   Manufacturability evaluation

Features extracted from the B-rep of the solid or obtained through the mapping pro described above are regions determined only by local topological and geometric cha teristics of the part. A visibility-based approach has been developed to find the gl attributes (accessibility) of features and determine how this affects the manufacturab of the part. An added gain from this approach has been the usefulness of the globa tributes in process planning. Local accessibility of a feature is first determined base the visibility of faces forming the feature and machining access directions are genera The global accessibility of each feature along the machining directions is then asse with respect to the part (Aditya Narayan 1995). Based on the global accessibility in mation it is possible to enumerate the machining sequences that will realise the part stock. This enumeration can then be pruned using other constraints such as dimensi
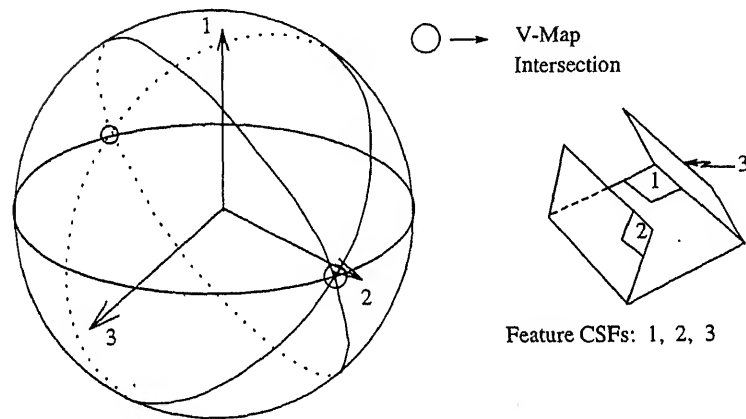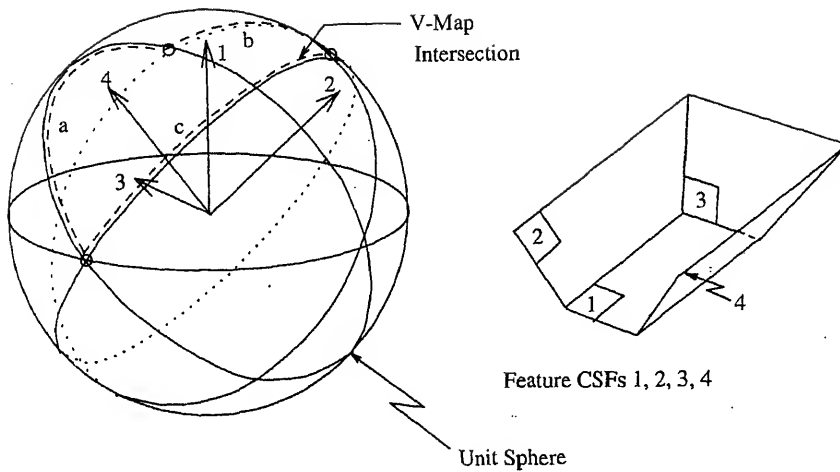
Figure 5. Visibility map (V_Map) of features.

references, set-up minimisation, tool availability etc. to generate machining process

cal accessibility of a feature refers to the existence of directions from which all points erest on the feature are visible. In the context of feature-based manufacturing, the tive is to find the directions from which all the faces of a feature are completely e. These are directions in which a tool might be able to approach a feature. sibility maps (V_Maps) (Woo & Chen 1992; Woo 1994) can be used to determine cal accessibility of a feature. A feature is said to be locally accessible if there is a tion in which all the created faces of a feature are visible. As a point in the V_Map urface denotes a direction from which all the points on the surface are visible, for e created faces of the feature to be visible, the V_Maps of the created faces must ect. The valid directions of access for the feature will lie within this intersection of V_Maps. This region of intersection of the V_Maps of the feature's created faces is

called the V_Map of the feature. Figure 5 shows the V_Map of two features. For the V-
shown on top, the V_Map is a finite region (indicated by the dotted line in figure 5). T
implies that there are many access directions for the feature. For the dovetail slot (bot
of figure 5) the intersection of the V_Maps of the created faces are just two points
are shown circled in the figure. The dovetail slot, therefore can be accessed for machir
only from the two directions that are normal to the two end faces.

Valid access directions are those which map to a point on the unit sphere that is abov
on the planes corresponding to the created faces of the feature. If no V_Map can be genera
for a feature, the feature is not locally accessible. Such features are to be handled either
changing the design of the feature to make them locally accessible or by resorting to spe
manufacturing methods, such as the use of special tools. Parts with such features are
completely manufacturable. At present such cases of local inaccessibility are flagge
the user.

The feature V_Map determines the local accessibility of features and is a characteristi
the feature only. They are sensitive only to orientation of the feature and not to its posi
in space. The feature V_Map is a semi-infinite set of directions, from which a finite se
directions has to be picked for evaluating global accessibility of the feature along th
directions. In actual machining practice, the choice of a particular access directio
based on heuristics. To generate as many relevant access directions as possible, candi
directions, based on some common cases encountered in general machining, are tes
Thus, normals to the created (CEFs and CSFs) and the shared faces (SEFs and SSFs)
checked with the feature V_Map and are taken as valid access directions if they lie withi
These directions form the initial set of access directions which will be checked for glc
accessibility of features.

7.1 *Global accessibility*

Determination of global accessibility is a check of whether a generic tool, representabl
a line, can reach the feature to machine it. Global accessibility of a feature does not ass
absolute manufacturability of the feature. It is only a first-step evaluation before a m
detailed tool and manufacturing environment specific accessibility analysis (that inclu
size effects) is done.

The traditional method of determining global accessibility by ray tracing from the reg
whose global accessibility is to be determined (Spyridi & Requicha 1990; Vandenbra
& Requicha 1993; Lim & Menq 1994) is a computationally intensive task if carried
for the whole feature. Further, ray tracing is susceptible to failure due to bad choice of
spacing and poor resolution. We define a feature to be globally accessible, along a directi
if the intersection of the part or workpiece and the extrusion of the feature volume in
direction is NULL i.e.

$$(iW \cap \text{Ext}(V_F)_l) = \Phi$$

where iW = interior of workpiece and $\text{Ext}(V_F)_l$ = extrusion to infinity of feature volu
V_F along a direction l (see figure 6). Mill· *et al* (1994) also use a volume-based appro
but their method does not yield correct results for directions of access that are not norr
to the assumed feature-machining face or for features such as dovetail slots.
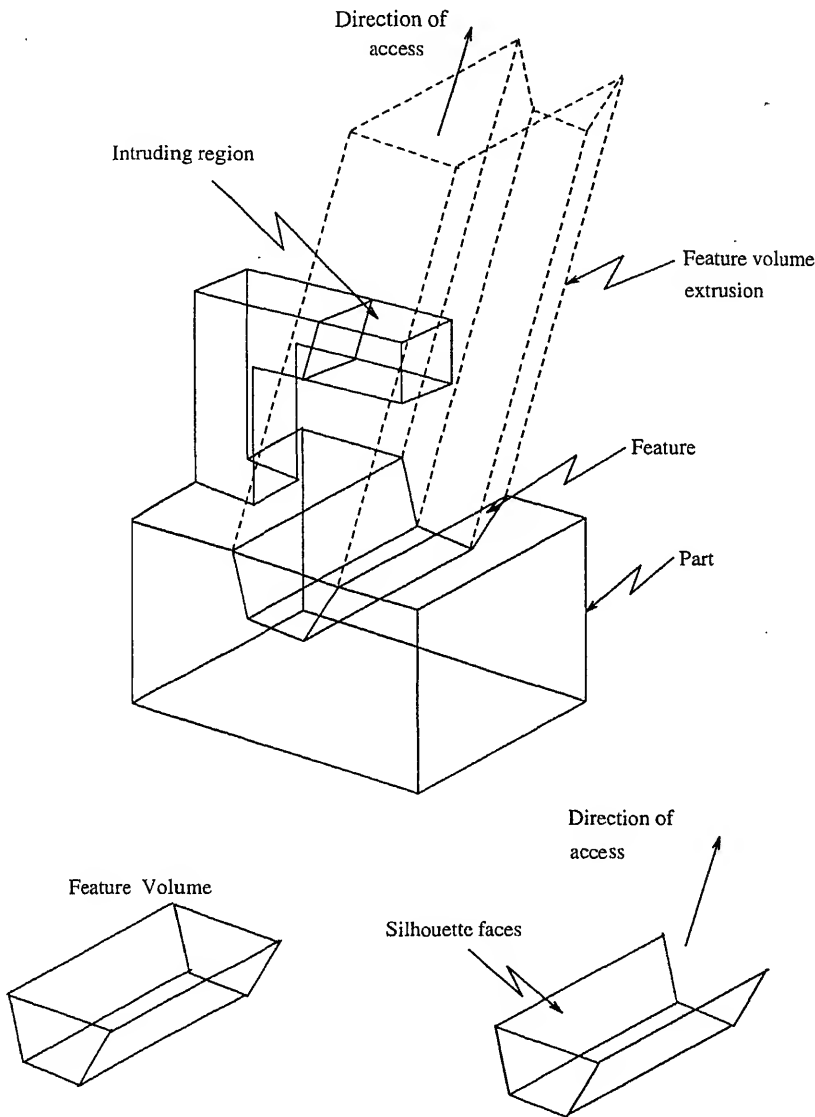
Figure 6. Determination of global accessibility of a feature.

The premise for using the feature volume to determine feature accessibility is that the ole feature volume, which has to be removed for creating a feature, should be accessible he tool. Figures 7 and 8 show the global accessibility of a feature along two directions. the direction in figure 7, the volume of intersection is NULL and hence the feature lobally accessible. For the direction in figure 8, it is not accessible, as a finite volume eturned. In both the figures, the first block shows the feature volume on the part and
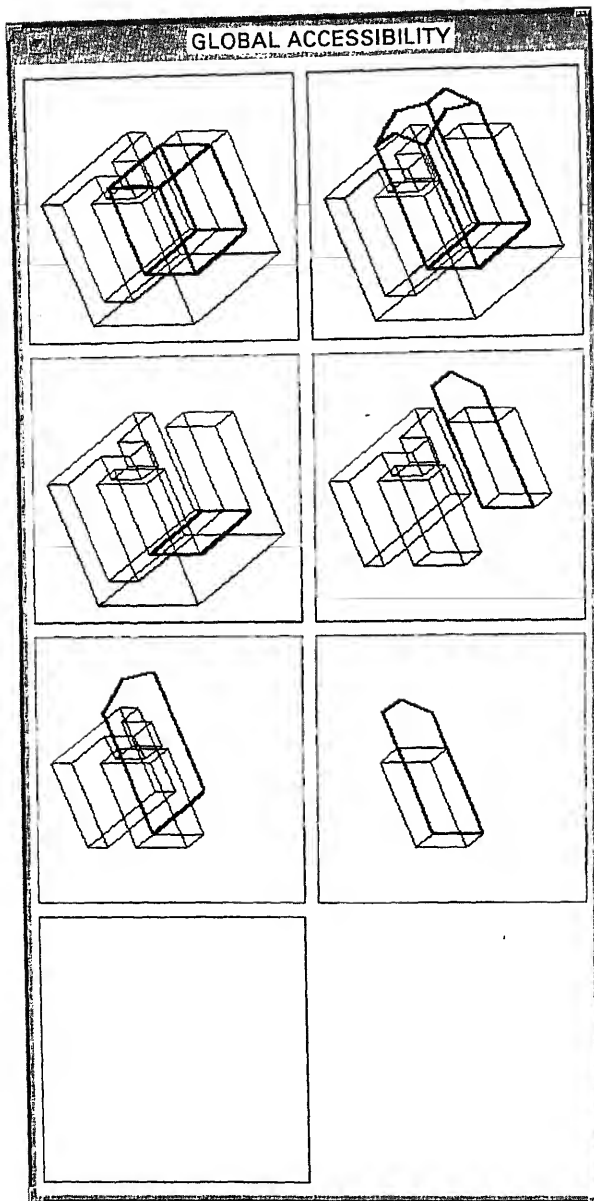
Figure 7. Globally accessible
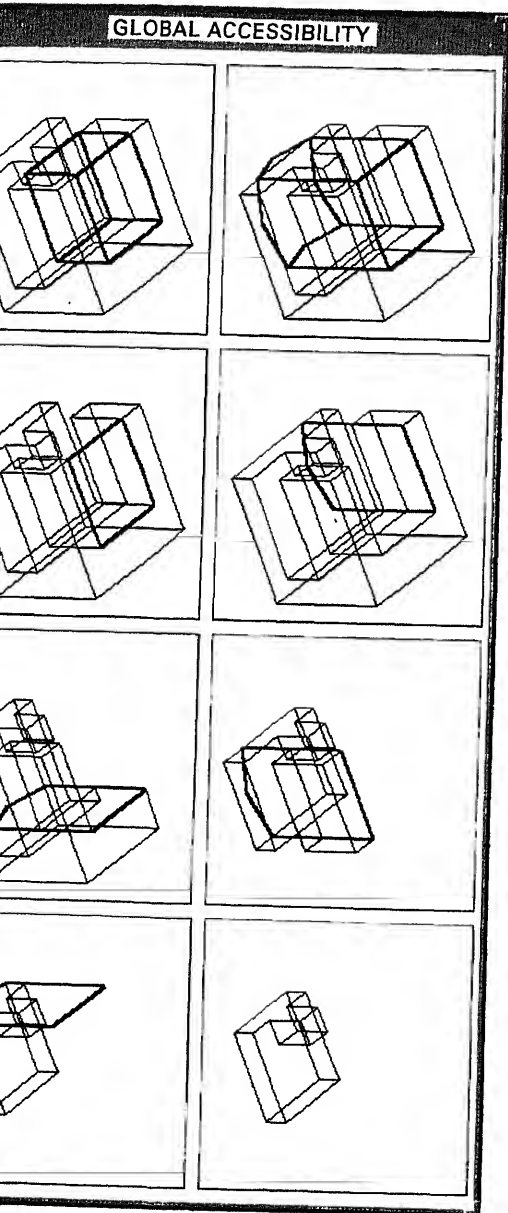
Figure 8. Globally inaccessible feature.

8.   Accessibility and process planning

Information about global accessibility of features in a part can be used to sequence
features for machining. If a feature is globally accessible with respect to the part, th
that feature can be created in the stock independent of other features (figure 9). Also, i
feature is not globally accessible then that feature can be created only after the creation
a feature from whose volume it is accessible. In figure 10, features B and C are globa
inaccessible features, whereas A is globally accessible. Features B and C can be crea
only if they are accessible with respect to a volume which has already been created. O
this precondition is satisfied they can be created in any sequence. Such features can
said to have *secondary accessibility* as they are dependent on creation of other features

Part (a)
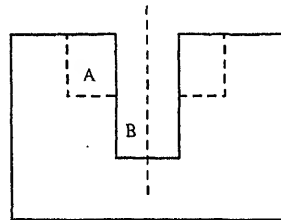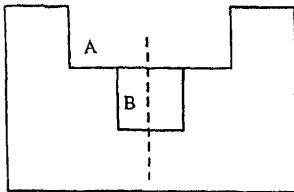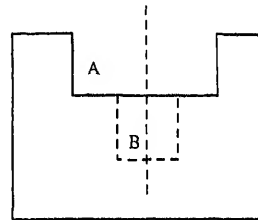
Interpretation 1   Interpretation 2

Part (b)
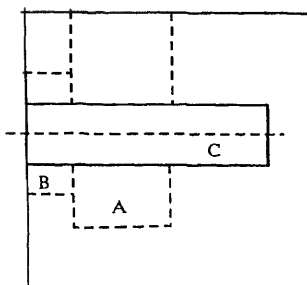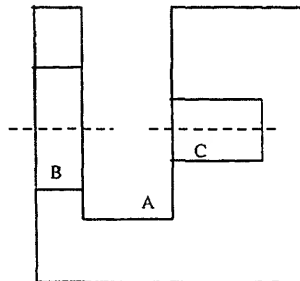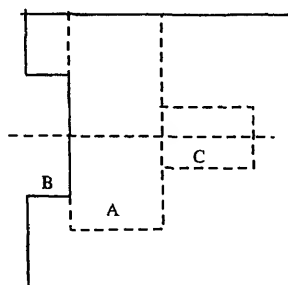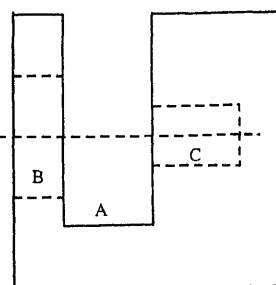
Interpretation 1   Interpretation 2   Interpretation 3
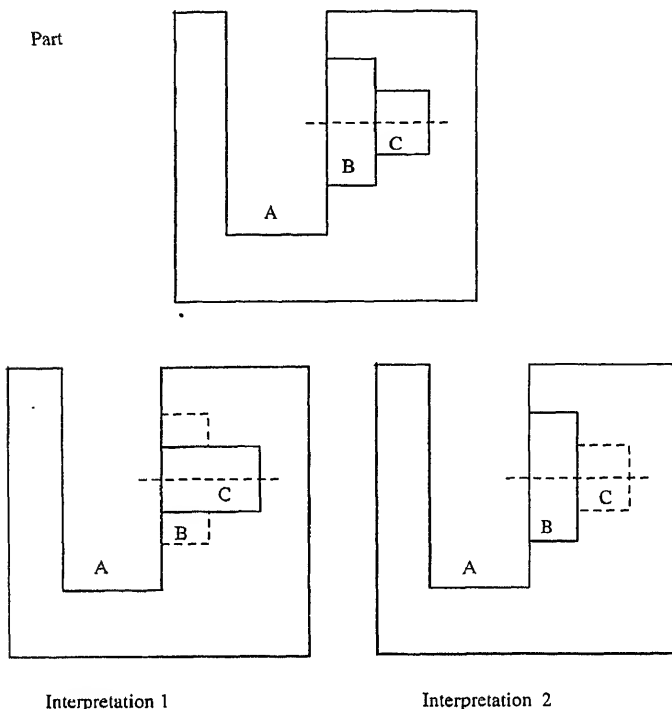
Figure 10. Machining sequence interpretations of globally inaccessible features.

becoming accessible. This dependency is not known *a priori* but has to be checked for the non-globally accessible features after the creation of every feature.

These assertions are based only on the consideration of spatial interrelationships between features and no other geometric or machinability-related constraints are operative. Based on the above, a tree of machining sequences, based on accessibility, can be generated where each path from the root to a leaf is a possible machining sequence from the accessibility point of view. At the root of the tree is the stock and at each leaf will be the final part if the part is manufacturable. At every level in the sequence tree, a feature will be created. The features at the first level are all globally accessible features as they can be created in the stock directly. At the other levels, the nodes will be the globally accessible features that have not been created at any node from the root to that node and features that gain secondary accessibility by the creation of features at ancestor nodes. An enumerated sequence tree is shown in figure 11 for a part with three features that are globally accessible and one feature that is globally inaccessible/exhibits secondary accessibility.

Features with secondary accessibility can be manufactured only if a tool–machine combination can be found that will reach it by fitting into an already created feature. These checks can be built into a tool selection module as constraints. This tree structure gives an exhaustive enumeration of all the possible machining sequences based on global accessibility. This allows the planner to consider all possible sequences in the search for an optimal process plan. There have been similar efforts reported (Nau *et al* 1992; Gupta & Nau 1993), that use similar volume-based checks. However, in these efforts, the features

Enumerated Sequence Tree

Figure 11.   Enumerated sequence tree for a sample part.

and the volumes associated with them are pre-defined. Feasible machining sequences
be generated from the tree by the application of machining domain constraints such
dimensions, set-up minimisation and machining optimization, to mention a few.

The authors have not encountered any earlier attempts at interpreting the dimensi
as constraints on machining sequence during generation of process plans. An attempt
been made to incorporate the dimensional information associated with the part design
process planning approaches (Aditya Narayan 1995).

The effect of specifying dimensional relationships between various entities in the p
such as abstract data, faces and edges, are mapped to features. The features are sequen
for machining based on these relationships. In the present implementation, only conv
tional linear and angular dimensions are handled. The dimension is assumed to be speci
between two entities of the part, either abstract (e.g. centerline of a hole) or real (e.g. f
edge). This is a bidirectional relationship in which either entity may be referenced from
other. No directionality of reference is assumed, though a feature or face may be 'forc
to be a reference face by not specifying another entity with it in the dimension struct

Figure 12. Test part with dimensions.

he sequence of machining based on dimensional data fulfills only the mandatory di-
ional conditions specified by the part design. Further optimization of the process may
rried out by considering the freedom offered in set–up generation, tool change, cost
nisation etc. Also, constraints on the process will be imposed by tool and machine tool
ility, their availability and fixturing. This can be done within the framework of the
l accessibility tree by evaluating the various sequences that are enumerated by the tree.
ce the machining features have been sequenced, the volumes corresponding to the
res can be found (Rao Nalluri *et al* 1995). Figure 13 shows the machining volumes
sponding to the creation of faces and features in each step of the sequence. Machining
res provide the logic for selection and sequencing of operations. Machining volumes
de the geometric information for generating NC tool paths. The NC path generation
le in figure 1 reasons over the machining volumes and generates the tool path (cutter
ion file) (Gurumoorthy 1996). In essence, once a component is modelled in the proto-
system, NC tool paths required to realise the part from a given stock are automatically
rated.

Concluding remarks

ave presented a framework based on features technology to support manufacturability
ation and process planning. A prototype system as shown in figures 1 and 2 is under
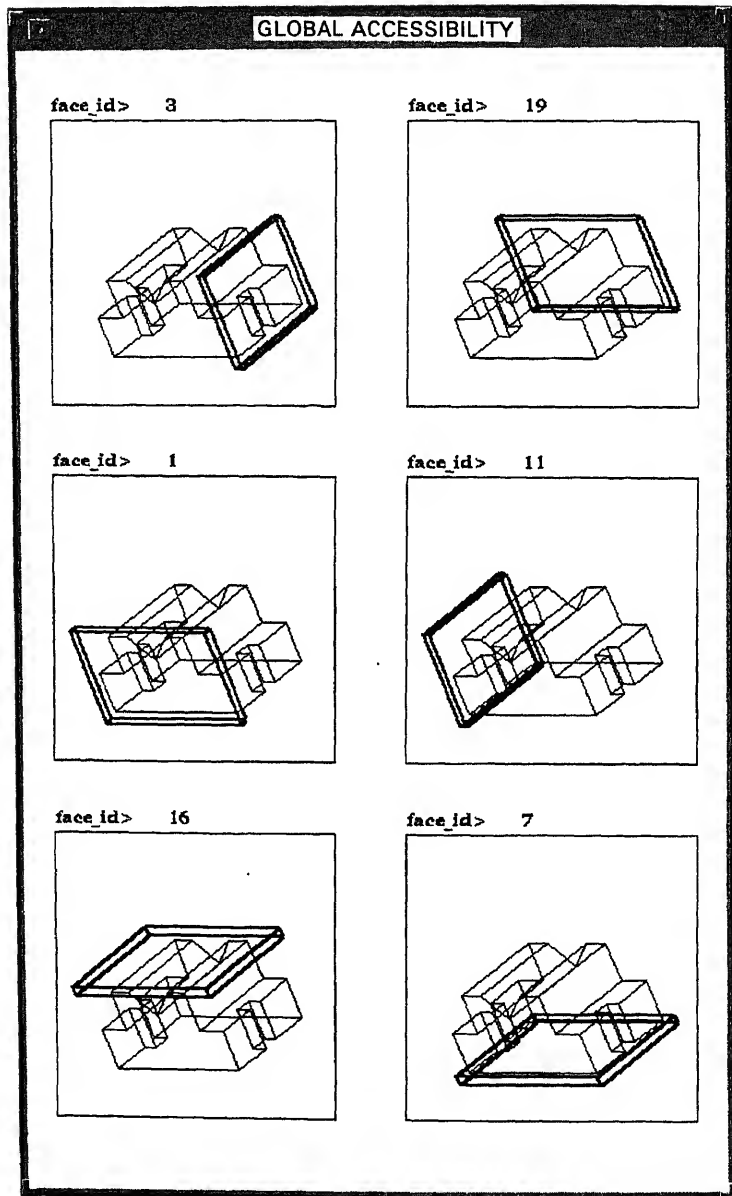
Figure 13.   Machining sequence.
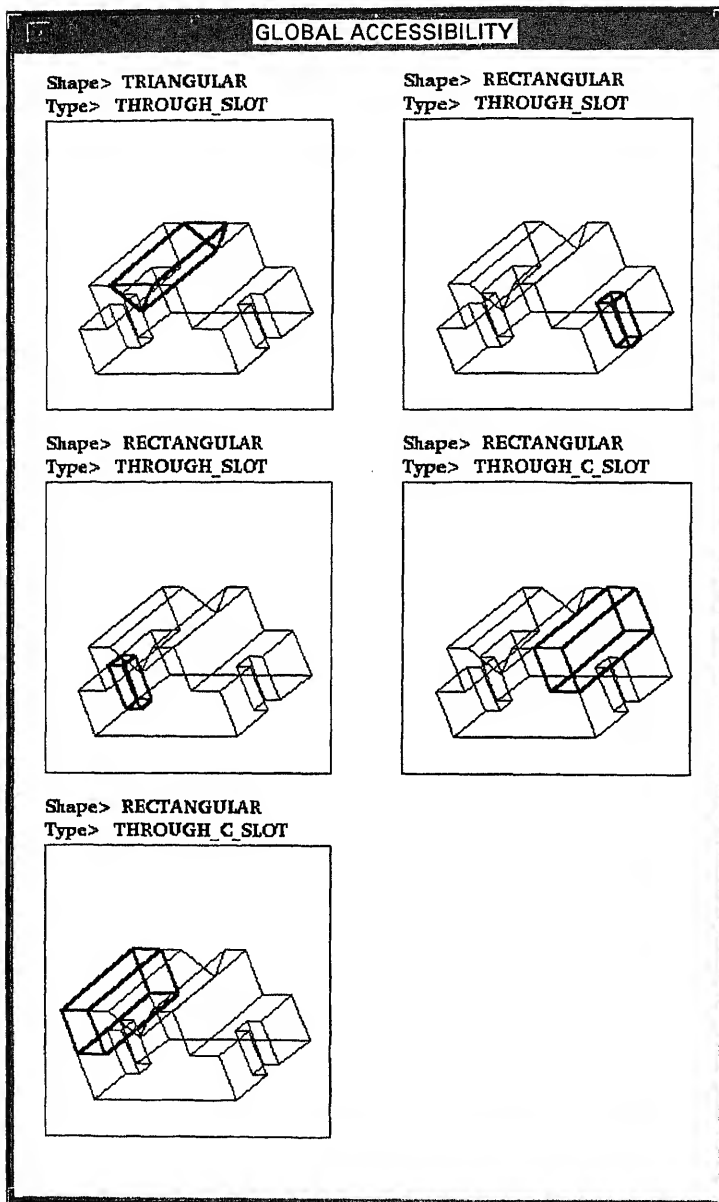
Figure 14. Machining sequence (contd.).

development. Feature-based modelling based on direct editing, feature extraction, feature mapping and generation of sequence tree based on global accessibility information and automatic tool path generation have already been implemented for polyhedral solids. Direct editing of feature-based models and feature extraction have also been implemented for exact B-Rep models using the ACIS geometric kernel (Hari 1995; Gupta 1997). Currently the focus is on implementing constraints based on good machining practice (as followed on the shop floor) into both the evaluation and sequencing tasks.

In this paper, limitations and bottlenecks in the feature technology have been brought out and remedies have been outlined. Defining form features through feature generation model is a novel approach which has significant potential for features technology. Separation of 'type' and 'shape' reduces the effort involved in implementation of feature extraction and feature based modelling. The enumeration of DIF feature types gives a sound basis for design and implementation of feature-based systems. By removing (through updating) the recognised features, part geometry as seen by the feature extraction module progressively simplifies and eliminates problems in feature extraction (interacting/intersecting features) that have been reported in the literature. Use of features has been extended to geometric manipulation through development of algorithms for direct manipulation and editing of geometric representation. Manipulation of product data in mapping requires further investigation.

It has been shown that issues such as global accessibility of features are vital not only for the evaluation of designs for manufacturability but also have an effect on the sequencing of machining operations. Currently, the accessibility analysis considers only the spatial relationships between features. For process planning, the effect of feature sizes is important for the evaluation of machine and tool capability and for optimising machining volume removal. Also, a complete tool accessibility check would require evaluating the exact shape and size of the tool and the feature. Algorithms to determine local and global accessibility have to be extended to handle the more general case of features with free-form surfaces. The current implementation considers dimensional data and references for sequencing of machining operations. Further work could focus on issues such as form tolerances, set-up minimisation, fixturing, tool and machine tool evaluation and machining practice and heuristics. The machining access directions determined based on the V_Map of features can be used for setup optimisation.

Global accessibility of features enable identifying multiple options in machining features. Work is underway to identify multiple feature interpretations across domains. For the integration of CAD geometric model with application domains to be completely automatic, redesign (i.e. modification of part geometry by applications) has to be possible. Some preliminary results are available in this area (Das *et al* 1996). The aim of the work presented in this paper has been to develop an algorithmic framework for process planning that is driven as much as possible by geometry and allows the incorporation of constraints and heuristics of the machining domain. It is hoped that further work in using DIF features as the unifying representation for geometric and domain specific reasoning will enable cor-
relation of domain knowledge and constraints with part geometry to support concurrent

ences

Narayan G 1995 *Process planning framework for feature based manufacturing*. Master
Eng. dissertation, Department of Mechanical Engineering, Indian Institute of Science,
galore

F 1992 *Requirements and architecture of CAM oriented CAD systems for design and
ufacture of mechanical parts*. Ph D thesis, University of California at Los Angeles,
Angeles, California

field W, Green M, Scott D, Stoker W 1986 Part features for process planning. CAM-I
ort, R-86-ppp-01, November

ian R A (ed.) 1992 *Design for manufacturability: Tool and manufacturing engineers hand-
k* (Dearborn, MI: Soc. Manuf. Eng.) vol 6

.997 Computer-aided Design Report. CAD/CAM Publishing, San Diego, CA

T C, Wysk R A 1985 *An introduction to automated process planning systems* (Englewood
ffs, NJ: Prentice Hall)

, Gupta S K, Nau D S 1996 Generating redesign suggestions to reduce setup cost: A step
ards automated redesign. *Comput. Aided Design* 28: 763–782.

S K 1997 *Feature-based editing of solid models*, ME dissertation, Department of Mechan-
Engineering, Indian Institute of Science, Bangalore

S K, Nau D S 1993 Generation of alternative feature-based models and precedence
erings for machining applications. In *Proceedings of ACM Solid Modeling Conference*,
ntreal

noorthy B 1996 Reverse engineering of surfaces and solids and automatic NC path genera-
1. Technical Report, TR/IISc/ME/CARL-P-96-3, Department of Mechanical Engineering,
ian Institute of Science, Bangalore

5 1995 Feature extraction from exact B-Rep models. ME dissertation, Department of
chanical Engineering, Indian Institute of Science, Bangalore

5, Chang T C 1988 Graph-based heuristics for recognition of machined features from 3D
d model. *Comput. Aided Design* 20(2): 58–66

nou L 1980 *Shape classification in computer aided design*. Ph D thesis, University of
nbridge, Cambridge, UK

P, Menq C H 1994 CMM feature accessibility and path generation. *Int. J. Product. Res.*
597–618

G, Naish J C, Salmon J C 1994 Design for machining for a simultaneous engineering
kstation. *Comput. Aided Design* 26: 521–527

S, Zhang G, Gupta S K 1992 Generation and evaluation of alternative operation sequences.
ality Assurance through Integration of Manufacturing Processes and Systems, ASME
nter Annual Meeting, pp 93–107

alluri S R P 1994 *Form features generation model for features technology*. Ph D dissertation,
partment of Mechanical Engineering. Indian Institute of Science, Bangalore

alluri S R P, Gurumoorthy B 1993a Knowledge-based gluing operators for feature-based
delling, *Comput. Ind.* 23: 129–138

alluri S R P, Gurumoorthy B 1993b Domain Independent Form Features for Concurrent
gineering. In *Proceedings of JSME-ASME Joint Workshop on Design*, Tokyo, pp 182–189

Rao Nalluri S R P, Vani V, Gurumoorthy B 1995 A 3D clipping algorithm for form feature volume extraction. In *Proceedings of International Conference on Computer Integrated Manufacturing*, Singapore, pp 385–393

Sakurai H, Dave P 1996 Volume decomposition and feature recognition. *Comput. Aided Design* 27: 519–537

Salomons O W, van Houten F J A M, Kals H J J 1993 Review of research in feature-based design. *J. Manuf. Syst.* 12: 113–132

Shah J J 1988 Feature transformations between application specific feature spaces. *Comput. Aided Eng. J.* December: 247–255

Shah J J 1991 Assessment of features technology. *Comput. Aided Design* 23: 331–343

Shah J J, Shen Y, Shirur A 1994 Determination of machining volumes from extensible sets of design features. In *Advances in feature based manufacturing* (eds) J J Shah, M Mantyla, D Nau (New York: Elsevier Science) pp 129–157

Spyridi A J, Requicha A A G 1990 Accessibility analysis for the automatic inspection of parts. In *Proceedings of IEEE International Conference on Robotics & Automation*, Cincinnati, pp 1284–1289

Subrahmanyam S, Wozny M 1995 Overview of automatic feature recognition techniques for computer-aided process planning. *Comput. Ind.* 26: 1–21

Tseng Y J, Joshi S B 1994 Recognizing multiple interpretations of interacting machining features. *Comput. Aided Design* 26: 667–688

Vandenbrande J H, Requicha A A G 1993 Spatial reasoning for the automatic recognition of machinable features in solid models. *IEEE Trans. Pattern Anal. Mach. Intell.* 15: 1269–1285

Woo T C, Chen L L 1992 Computational geometry on the sphere with application to automated machining. *Trans. ASME J. Mech. Design* 114: 288–295

Woo T C 1994 Visibility maps and spherical algorithms. *Comput. Aided Design* 26: 6–16

optimization-based algorithm for job shop scheduling

JIHUA WANG, PETER B LUH, XING ZHAO and JINLIN WANG

Department of Electrical and Systems Engineering, University of Connecticut, Storrs, CT 06269-2157, USA
e-mail: Luh@brc.uconn.edu

Abstract. Scheduling is a key factor for manufacturing productivity. Effective scheduling can improve on-time delivery, reduce inventory, cut lead times, and improve the utilization of bottleneck resources. Because of the combinatorial nature of scheduling problems, it is often difficult to find optimal schedules, especially within a limited amount of computation time. Production schedules therefore are usually generated by using heuristics in practice. However, it is very difficult to evaluate the quality of these schedules, and the consistency of performance may also be an issue.

In this paper, near-optimal solution methodologies for job shop scheduling are examined. The problem is formulated as integer optimization with a "separable" structure. The requirement of on-time delivery and low work-in-process inventory is modelled as a goal to minimize a weighted part tardiness and earliness penalty function. Lagrangian relaxation is used to decompose the problem into individual part subproblems with intuitive appeal. By iteratively solving these subproblems and updating the Lagrangian multipliers at the high level, near-optimal schedules are obtained with a lower bound provided as a byproduct. This paper reviews a few selected methods for solving subproblems and for updating multipliers. Based on the insights obtained, a new algorithm is presented that combines backward dynamic programming for solving low level subproblems and interleaved conjugate gradient method for solving the high level problem. The new method significantly improves algorithm convergence and solution quality. Numerical testing shows that the method is practical for job shop scheduling in industries.

Keywords. Scheduling; Lagrangian relaxation; dynamic programming.

Introduction

eduling is a key factor for manufacturing productivity. Effective scheduling can im-
ve on-time delivery, reduce inventory, cut lead times, and improve the utilization of
leneck resources. Because of the combinatorial nature of scheduling problems, it is of-
difficult to obtain optimal schedules, especially within a limited amount of computation

time. Production schedules therefore are usually generated by using heuristics in practice. However, it is very difficult to evaluate the quality of these schedules, and the consistency of performance may also be an issue. A logical strategy is thus to pursue methods that can consistently generate good schedules with quantifiable quality in a computationally efficient manner.

This paper examines the practical scheduling of job shops, a typical environment for the manufacture of low-volume and high-variety parts. In a job shop, parts with various due dates and priorities are to be processed on various types of machines. Job shop scheduling is to select the machines and beginning times for individual operations to achieve certain objective(s) with given machine capacities. In this paper, job shop scheduling is formulated as integer optimization with a "separable" structure. The requirement of on-time delivery and low work-in-process inventory is modelled as a goal to minimize a weighted part tardiness and earliness penalty function. Lagrangian relaxation (LR) is used to decompose the problem into individual part subproblems with intuitive appeal. By iteratively solving those subproblems and updating the Lagrangian multipliers at the high level, near-optimal schedules are obtained with a lower bound provided as a byproduct on the optimal cost. This paper reviews a few selected methods for solving subproblems and for updating multipliers. Based on the insights obtained, a new algorithm is presented that combines "backward" dynamic programming (BDP) for solving low level subproblems and interleaved conjugate gradient (ICG) method for solving the high level problem. The new method significantly improves algorithm convergence and solution quality. Numerical testing shows that the method is practical for job shop scheduling in industries.

1.1   *Literature review*

Given the economic and logistical importance of the scheduling problem, many of the early efforts centred on obtaining optimal schedules. Two prominent optimization methods are the branch and bound method (Fisher 1973) and dynamic programming (e.g., Pinedo 1995). It was discovered that the generation of optimal schedules often requires excessive computation time regardless the methodology. Furthermore, job shop scheduling is among the hardest combinatorial optimization problems and is NP-complete (Garey & Johnson 1979). Production schedules therefore are usually generated by experienced shop-floor personnel using simple dispatching rules in practice. Many heuristic methods have been presented and implemented based on due dates, criticality of operations, operation processing times, and machine utilization (e.g., Blackstone *et al* 1982). Many artificial intelligence (AI) approaches also use heuristics for scheduling (e.g., Kuziak 1990). These heuristics-based approaches usually generate feasible schedules quickly but it is very difficult to evaluate the quality of the schedules. Also, most heuristics do not provide for iterative improvement of the schedules.

Attempts to bridge the gap between heuristic and optimization approaches have also been undertaken (Adam *et al* 1988; Luh & Hoitomt 1993; Ventura & Weng 1995). In Adams *et al* (1988), for example, a heuristic for job shop scheduling was developed based upon

he solution was fed back into the previously solved machine problem by a "local imization." However, schedule evaluation could only be achieved through "selective neration." Also, each operation has to be pre-assigned to a specific machine before luling, though the operation may be processed on different machines or different types achines.

Luh & Hoitomt (1993), a Lagrangian relaxation framework was established for manu-ring system scheduling problems, and a practical method was provided. In the method, machine capacity and operation precedence constraints are relaxed by using Lagrange ipliers, and operation-level subproblems are formed and solved by enumeration. The ipliers are then updated at the high level by using a subgradient method. An improved on of the method considering bills of materials, with a modified subgradient method e high level was presented in Czerwinski & Luh (1994).

uch progress has been made on the scope and performance of the LR-based method-y. A combined LR and heuristic method was developed for job shop scheduling with p-dependent setups and finite buffers in Luh *et al* (1995). The scheduling of batch nines with setup requirements was addressed in Luh *et al* (1997b). A "forward" dy-ic programming (FDP) algorithm was embedded within the LR framework for job scheduling in Chen *et al* (1995). In the method, only machine capacity constraints elaxed, and part level subproblems are formed and solved by using the FDP. By doing the solution oscillation difficulties as reported in Czerwinski & Luh (1994) are alle-ed. Also, by relaxing less constraints, the dual cost should be a tighter lower bound. uh *et al* (1997), an LR-based method was developed for job shop scheduling with ertain parts. A "backward" dynamic programming (BDP) was developed to solve the subproblems with random part parameters.

or high level algorithms, the slow convergence of subgradient methods was analysed, the facet ascending algorithm (FAA) was presented to improve the convergence in astik & Luh (1993). The reduced-complexity bundle method (RCBM) was developed omastik & Luh (1996) which significantly reduces the computation complexity but itains the convergence of the conventional bundle methods. An interleaved subgradient nod (ISG) was developed in Kaskavelis & Caramanis (1995) to improve the efficiency ie LR-based method. A review of those methods will be presented in § 3.3.

Overview of the paper

2, an integer optimization formulation with a "separable" structure for job shop duling is presented. In § 3, the problem is decomposed into individual part subprob-s by relaxing machine capacity constraints following the approach in Chen *et al* (1995). w selected methods for solving subproblems and for updating multipliers are reviewed. ed on the insights obtained, a new algorithm is presented that combines BDP for solv-low level subproblems and an interleaved conjugate gradient (ICG) method for solving high level problem. In § 4, numerical results show that the new method outperforms a ious LR/SG method in convergence. Numerical testing for practical data sets shows the method can generate high quality schedules in a timely fashion.

2. Problem formulation

In a job shop, machines may have different processing capabilities. Machines with the same processing capability are grouped as a "machine type," and all the machine type form a set denoted as H. The total number of machine types is thus $|H|$. There are I parts with various due dates D_i to be scheduled over a discretized time horizon K. Part i ($i = 0, 1, \ldots, I - 1$) consists of nonpreemptive J_i serial operations with operation j ($j = 0, 1, \ldots, J_i - 1$) of part i denoted by (i, j). An operation may·start only after its preceding operation has been completed, and requires a machine belonging to a given set of eligible machine types H_{ij} for a specified duration of time. Without loss of generality, it is assumed that operations of part i are performed in the ascending order of operation index j.

The time horizon consists of K time units, indexed by k ($k = 0, 1, \ldots, K - 1$). Each operation beginning time is defined as the beginning of the corresponding time unit, and each completion time the end of the time unit. The variables used in the problem formulation are listed below.

δ_{ijhk}:   0-1 operation variable which is one if operation (i, j) is performed on machine type h at time k, and zero other.

β_i: Part earliness weight.

b_{ij}:   Beginning time of operation (i, j).

c_{ij}:   Completion time of operation (i, j).

D_i: Due date of part i.

E_i: Earliness of part i, defined as $E_i = \max[0, S_i - b_{i0}]$.

h: Machine type variable, $h \in H$.

H_{ij}:   Set of machine types capable of performing operation (i, j).

J: Objective function to be minimized.

k: Time index ($k = 0, 1, \ldots, K - 1$).

M_{hk}:   Capacity of machine type h at time k.

P_{ijh}:   Processing time of operation (i, j) on machine type $h \in H_{ij}$.

S_i: Desired raw material release time for part i.

T_i: Tardiness of part i, defined as $T_i = \max[0, c_{i,J_i-1} - D_i]$.

W_i: Part tardiness weight.

Assuming that the set of machine types, the number of parts, part due dates and weights, operation processing time and time horizon are given, the constraints and objective function are explained below.

2.1 Machine capacity constraints

In the literature (e.g., Baker 1974; Adams 1988), the limited machine capacity is often modelled by "disjunctive constraints." Given a pair of operations (denoted as A and B) to be performed on a particular machine, the disjunctive constraints state that *either B starts after the completion of A or A starts after the completion of B*. As a result, the number of disjunctive constraints increase drastically with the number of operations. Also, it is

ed that each operation must be pre-assigned to specific machines, though an operation
be processed on different machines or different types of machines.

defining a set of 0-1 operation variables δ_{ijhk} to represent the processing status of
operation, the following machine capacity constraints are formed. The constraints
that the total number of operations being performed (active) on machine type h must
s than or equal to the capacity (M_{hk}) of machine type h at any time unit k, i.e.,

$$\sum_{i=0}^{I-1} \sum_{j=0}^{J_i-1} \delta_{ijhk} \leq M_{hk}, \quad h \in H; \quad k = 0, \dots, K-1, \tag{1}$$

e the 0-1 operation variable δ_{ijhk} is defined by

$$\delta_{ijhk} = \begin{cases} 1, & \text{if } b_{ij} \leq k \leq c_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

e number of machine capacity constraints equals the number of machine types times
me horizon K. Although the number of 0-1 operation variables is huge, these variables
etermined once the machine types and beginning times of the operations are speci-
They thus are not independent decision variables, and do not cause any complexity
ulty.

Operation precedence constraints

peration precedence constraints are represented by the following "conjunctive con-
ts." These constraints state that an operation cannot be started until its preceding
tion is finished, i.e.,

$$c_{i,j-1} + 1 \leq b_{ij}, \quad i = 0, 1, \dots, I-1; \quad j = 1, 2, \dots, J_i - 1. \tag{3}$$

ice operation $(i, j-1)$ is completed at the end of time unit $c_{i,j-1}$, and operation (i, j)
at the beginning of time unit b_{ij}, the term "1" is required in (3). For the same reason,
rm "1" also appears in the following processing time requirements.

Processing time requirements

processing time requirements state that each operation must be assigned the required
int of time for processing on the selected machine type h, i.e.,

$$c_{ij} = b_{ij} + P_{ijh} - 1, \quad i = 0, 1, \dots, I-1; \quad j = 0, 1, \dots, J_i - 1; \quad h \in H_{ij}. \tag{4}$$

ith processing times specified, operation completion times c_{ij} can be eliminated from
roblem formulation. For notational convenience, they still appear in later derivation.

Objective function

bus objective functions such as makespan have been used in the literature. Research

useful than, say, makespan criteria (Blackstone *et al* 1982). In addition, the additivity of the tardiness objective function facilitates the decomposition approach.

Besides the on-time delivery, working-in-process (WIP) inventory is another major concern in practice. To reduce WIP inventory, a desired raw material release time S_i for each part is derived based on part due date and total processing time of the part (sum of operation processing times of the part). An earliness term for each part is added to the tardiness objective function, representing the penalty for releasing raw material too early. The requirement for on-time delivery and low WIP inventory is thus modelled as a goal to minimize the weighted part tardiness and earliness penalties, i.e.,

$$J \equiv \sum_{i=0}^{I-1} (W_i T_i^2 + \beta_i E_i^2). \tag{5}$$

The square on tardiness reflect the fact that a part becomes more critical with each time unit after passing its due date. Similarly, square is applied to each earliness penalty term. The objective function accounts for the priorities of the parts, the importance of meeting due dates and desired release times.

The overall problem therefore is to minimize the part tardiness and earliness penalty function, subject to the above machine capacity and operation precedence constraints, i.e.,

$$\min_{\{b_{ij}, h_{ij}\}} J, \text{ with } J \equiv \sum_{i=0}^{I-1} (W_i T_i^2 + \beta_i E_i^2), \tag{6}$$

subject to

$$\sum_{i=0}^{I-1} \sum_{j=0}^{J_i-1} \delta_{ijhk} \leq M_{hk}, \quad h \in H; \quad k = 0, \ldots, K-1, \tag{7}$$

$$c_{i,j-1} + 1 \leq b_{ij}, \quad i = 0, 1, \ldots, I-1; \quad j = 1, 2, \ldots, J_i - 1. \tag{8}$$

The decision variables are the operation beginning times b_{ij} and the machine types h_{ij} for individual operations. Once b_{ij} and h_{ij} are selected, $\{c_{ij}\}$, $\{T_i\}$, $\{E_i\}$, and $\{\delta_{ijhk}\}$ can be easily derived.

3. Solution methodology

3.1 *Lagrangian relaxation*

Lagrangian relaxation (LR) is a mathematical programming technique for performing constrained optimization. Similar to pricing concept of a market economy, the Lagrangian relaxation method replaces "hard" coupling constraints (e.g., machine capacity constraints) by the payment of certain "prices" (i.e., Lagrange multipliers) for the use of machines at individual time units. The original NP-hard problem can thus be decomposed into many smaller and easier subproblems. The solutions of individual subproblems, when put together, may not constitute a feasible schedule since coupling constraints have been relaxed by the multipliers. These prices or multipliers are thus iteratively adjusted based on the degree of constraint violations following again the market economy mechanism.

blems are then re-solved based on the new set of multipliers. In mathematical terms, function" is maximized in this multiplier updating process, and values of the dual n serve as lower bounds to the optimal feasible cost. At the termination of this lier updating process, simple heuristics are used to adjust subproblem solutions to e a feasible schedule satisfying all constraints. Heuristics can also be run after each zation iteration to check convergence or to provide candidate feasible schedules. ization and heuristics thus operate in a synergistic fashion to generate effective les. The quality of the schedule can also be quantitatively evaluated by comparing t to the largest lower bound provided by the dual function.

using Lagrange multipliers π_{hk} to relax machine capacity constraints, the following d problem is obtained.

d problem

$$
\min_{\{b_{ij}, h_{ij}\}} L, \text{ with } L \equiv \sum_i (W_i T_i^2 + \beta_i E_i^2) + \sum_{i,j} \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k} - \sum_{h,k} M_{hk} \pi_{hk}, \tag{9}
$$

t to the operation precedence constraints (3).

leriving (9), the fact $\sum_{h \in H} \sum_{k=0}^{K-1} \pi_{hk} \delta_{ijhk} = \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k}$ is used, and the relaxed em has b_{ij} and h_{ij} as its decision variables. After regrouping terms related to indi- parts, the relaxed problem can be decomposed into the following part subproblems.

ubproblems

$$
\min_{\{b_{ij}, h_{ij}\}} L_i, \text{ with } L_i \equiv W_i T_i^2 + \beta_i E_i^2 + \sum_{j=0}^{J_i-1} \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k}, \tag{10}
$$

t to the corresponding operation precedence constraints for part i.

m (10), a part subproblem reflects the needs to balance tardiness penalty, earliness y, and machine utilization costs. This part subproblem can be viewed as a multi-stage ization problem with each stage corresponding to an operation. Although solving the al problem by using dynamic programming (DP) is impractical, the decomposed part oblem is not NP-hard, and can be efficiently solved by using DP as will be presented .2.

L_i^* denote the minimal subproblem cost of part i with given multipliers, the high Lagrangian dual problem is then obtained as below.

problem

$$
\max_{\{\pi_{hk}\}} D, \text{ with } D \equiv \sum_i L_i^* - \sum_{h,k} M_{hk} \pi_{hk}. \tag{11}
$$

The Lagrangian dual function D is concave (Bertsekas 1995), and piece-wise linear, and consists of many "facets" (Tomastik & Luh 1993). We next present the resolution of part subproblems, followed by the updating of Lagrange multipliers.

3.2 Dynamic programming

The forward dynamic programming (FDP) algorithm presented by Chen *et al* (1995) can be used to solve a part subproblem in (10). It starts with the first operation of the part, and precedes to the last operation. In this paper, a backward dynamic programming (BDP) is developed with the goal to be further extended to handle uncertainties (e.g., uncertain arrival times, processing times, due dates etc., see Luh *et al* 1997). The BDP algorithm starts with the last stage, and compute the costs of the last operation $(i, J_i - 1)$ for all possible b_{i,J_i-1} and h_{i,J_i-1}:

$$V_{i,J_i-1}(b_{i,J_i-1}, h_{i,J_i-1}) = W_i T_i^2 + \sum_{k=b_{i,J_i-1}}^{c_{i,J_i-1}} \pi_k h_{i,J_i-1}. \tag{12}$$

For other operations (i, j), the cumulative costs are obtained by recursively solving the following DP equation subject to operation precedence constraints (3):

$$V_{ij}(b_{ij}, h_{ij}) = \min_{b_{i,j+1}, h_{i,j+1}} \left\{ \Delta_{ij} \beta_i E_i^2 + \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k} + V_{i,j+1}(b_{i,j+1}, h_{i,j+1}) \right\}$$

$$= \Delta_{ij} \beta_i E_i^2 + \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k} + \min_{b_{i,j+1}, h_{i,j+1}} V_{i,j+1}(b_{i,j+1}, h_{i,j+1}). \tag{13}$$

In the above, Δ_{ij} is 1 if (i, j) is the first operation $(j = 0)$ and 0 otherwise. The function $V_{ij}(b_{ij}, h_{ij})$ is the cumulative cost for all operations succeeding and including (i, j), and $\Delta_{ij} \beta_i E_i^2 + \sum_{k=b_{ij}}^{c_{ij}} \pi_{h_{ij}k}$ are the "stage-wise" costs. The algorithm starts from the last stage and moves backwards till the first stage is reached. The optimal subproblem cost L_i^* is then obtained as the minimal cumulative cost at the first stage. Finally, the optimal beginning times b_{ij} and machine types selected h_{ij} for operations can be obtained by tracing forwards the stages. Similar to FDP, the computation complexity of the above BDP algorithm is $O(K \sum_{j=1}^{J_i} |H_{ij}|)$ (Luh *et al* 1997).

3.3 Solving dual problem

3.3a *Subgradient methods*: As mentioned above, the Lagrangian dual function is concave and piece-wise linear. Existing methods for optimizing the dual function fall roughly into three classes: subgradient, cutting plane, and bundle methods. Of these, subgradient methods are commonly used to update the Lagrange multipliers (i.e., to maximize the dual function) because of their simplicity, the speed for computing a direction, and the global convergence property. With the subproblem solutions for given multipliers π_{hk}, the subgradient g of the dual function D is calculated by

$$g_{hk} = \sum_{i=0}^{I-1} \sum_{j=0}^{J_i-1} \delta_{ijhk} - M_{hk}, \quad h \in H; \quad k = 0, \ldots, K-1, \tag{14}$$

where g_{hk} is an element of the subgradient. In subgradient methods, multipliers are updated along the direction of the subgradient with the step size determined by

$$\alpha^n = \gamma \frac{D^* - D^n}{(g^n)^T g^n}, \qquad 0 < \gamma \le 2 \tag{15}$$

where D^* is the optimal dual cost, and α^n, D^n and g^n are respectively the step size, dual cost and subgradient at iteration n. As shown in Tomastik & Luh (1993), subgradient methods often zigzag across a ridge (intersection of some facets) of the dual function. The slow convergence rate (less than linear) of the subgradient methods causes these methods to require many iterations to reach an optimum.

3.3b *Facet ascending algorithm*: By recognizing that the Lagrangian dual function is polyhedral concave, and is made up of many facets, the facet ascending algorithm (FAA) finds the intersection of adjacent facets (Tomastik & Luh 1993). A subgradient of one of the facets is then projected on the intersection to obtain an ascending direction, and a line search technique is used to determine how far to move along the direction. The FAA avoids the zigzagging behaviour of subgradient methods, and shows improved convergence. For large problems, an intersection is usually formed by many facets. Finding such an intersection is often difficult and requires many dual function evaluations which are "computationally expensive." Furthermore, the ridges are short, causing slow convergence.

3.3c *Bundle methods*: The bundle method (e.g., Hiriart-Urruty & Lemarechal 1993) has the fastest convergence rate among the three classes of methods. It accumulates and utilizes the subgradients of points within a neighbourhood of the current iterate to find an ϵ-ascent direction (along which the function value can increase at least by ϵ), or to detect within ϵ of the dual optimum (ϵ-optimal). Finding such a direction or detecting ϵ-optimal, however, requires solving a number of quadratic programming problems with considerable complexity. To reduce the complexity while maintaining the convergence of the bundle method, the reduced-complexity bundle method (RCBM) finds an ϵ-ascent direction by performing a projection of a subgradient onto an appropriate subspace formed by the subgradients in the bundle (Tomastik & Luh 1996). Along the ϵ-ascent direction, a line search technique is then used to determine the step size for updating multipliers. Similar to FAA, the large number of dual function evaluations required to accumulate the subgradients and to perform line search is very time consuming, and hinders the applicability of the RCBM to very large problems.

3.3d *Interleaved subgradient method*: The iterative resolution of the dual problem requires the dual function to be evaluated many times, and each function evaluation involves solving all the subproblems once (called one iteration). These dual function evaluations are extremely "expensive" for large problems. For example, it takes about 68% of total CPU time for a case with 82 parts and 14 machines. To efficiently utilize the expensive function

evaluations, an interleaved subgradient (ISG) method has been developed (Kaskavelis & Caramanis 1995). Instead of solving all subproblems before updating multipliers, the ISG method updates multipliers after solving each subproblem. At the high level, the multipliers are updated along the direction of the subgradient. Numerical results show that the ISG method converges much faster than a subgradient method, though algorithm convergence has not yet been established.

3.3e Interleaved conjugate gradient method:

3.3e   Interleaved conjugate gradient method: As mentioned earlier, the dual function is concave, piece-wise linear, and consists of many facets. Each possible solution of the relaxed problem corresponds to a facet. Because of the combinatorial nature of the original problem, the number of possible solutions of the relaxed problem and therefore the number of facets increases drastically as the problem size increases. The dual function thus approaches a smooth function. This "smoothness" of the dual function motivates the use of optimization methods for smooth functions.

Among the methods for optimizing smooth functions, conjugate gradient methods have attractive convergence properties and computation efficiency. The conjugate directions are generated by

$$d^n = g^n + \beta^n d^{n-1} \text{ with } d^0 = g^0, \quad \beta^n = \frac{(g^n)^T g^n}{(g^{n-1})^T g^{n-1}}, \quad n = 1, 2, \dots.$$

(16)

where d^n and g^n are the conjugate direction and gradient at iteration n. The step size for updating multipliers is determined by performing a line search along the conjugate direction.

By incorporating the "interleave" concept with the conjugate gradient method, an interleaved conjugate gradient (ICG) method has been developed that utilizes the "smooth" property of the dual function and efficiency of the interleaved method for problems of large sizes (e.g., 1000 multipliers or more). In this paper, the ICG method is used to update the multipliers. The ICG algorithm is summarized as follows.

S0 Given the initial multipliers, solve all the part subproblems, and compute the dual cost and subgradient. Update multipliers along the direction of the subgradient. Set subproblem index $s = 1$.

S1 Solve subproblem s while keeping other subproblem solutions unchanged. Compute the "surrogate" dual cost and subgradient according to (11) and (14) with the latest available subproblem solutions.

S2 Compute conjugate direction by (16) with the surrogate subgradient, and update multipliers along the direction. Since only one subproblem is solved for a set of multipliers, line search cannot be used to determined the step size. The step size is therefore still computed according to (15) with D^* replaced by the lowest feasible cost obtained up to the current iteration.

S3 Increase s by one. If s is larger than the total number of subproblems, reset s to 1. Go to S1.

3.4   *Constructing feasible schedule*

The solutions to part subproblems, when put together, are generally associated with an infeasible schedule, i.e., capacity constraints might be violated at some time periods. A feasible schedule is constructed by using a list scheduling heuristic. In the list scheduling procedure, a list of immediately performable operations is created, and maintained in the ascending order of their beginning times from part subproblem solutions. Operations are then scheduled on the required machine types according to this list as machines become available. If the capacity constraint for a particular machine type is violated at time k, a greedy heuristic determines which operations should begin at that time and which ones are to be delayed by one time unit. The subsequent operations of those delayed ones are then delayed by one time unit if precedence constraints are violated. The process repeats until the last operation in the list.

The cost of the feasible schedule J is an upper bound on the optimal cost J^*. The optimal dual value D^*, on the other hand, is a lower bound on J^*. Since it is usually difficult to find J^* and D^*, the (relative) duality gap $(J - D)/D$ is often used as a measure of the quality of the feasible schedule.

4.   Numerical results

The new method that combines BDP and ICG within the LR framework has been implemented using the object-oriented programming language $C++$, and extensive testing has been performed. Four test cases are presented below. The first two cases are to demonstrate that the new LR/BDP/ICG method has better convergence than the previous LR/SG method with both machine capacity and operation precedence constraints relaxed and with a subgradient method at the high level. Case 2 also shows that the LR/BDP/ICG method generates a better schedule than a heuristic method that combines the "first come first serve (FCFS)" and "shortest processing time (SPT)" rules (called FCFS/SPT). The next two cases demonstrate that the LR/BDP/ICG method is applicable for solving scheduling problems of realistic sizes. In presenting the results for both LR/SG and LR/BDP/ICG methods, an iteration corresponds to solving all subproblems once.

The four cases are tested on a Sun Sparc 10 workstation. In the testing, all multipliers are initialized at zero. The time horizons are automatically generated based on machine availabilities and part processing requirements. The step size factor γ in the LR/BDP/ICG method is initialized to a specific value, and adaptively adjusted based on information obtained in the iterative process.

Case 1:   This test case is to demonstrate that the LR/BDP/ICG method generates a tighter lower bound than the LR/SG method. There are two machines of different types and two parts. Part one has two operations with processing times 3 and 2, respectively, and part two also has two operations with processing times 1 and 4, respectively. The first operations of both parts require machine type 0, and the second operations require machine type 1. The due dates are zero. The tardiness penalty weights are 1, and there is no earliness penalty. The results are summarized in table 1.

Table 1. Testing results for case 1 ($\gamma = 0.5$).

Method	Iteration	Dual/J	Duality gap	CPU (s)
LR/SG	2000	44/52	18%	3.00
LR/BDP/ICG	17	52/52	0%	0.01

It can be seen that both methods generate optimal schedule with the cost $J = 52$ which equals to the lower bound obtained by LR/BDP/ICG. The dual cost by using LR/BDP/ICG is thus optimal. The dual cost obtained by LR/SG ($= 44$) is also optimal for the corresponding dual function, as verified by using a LR/RCBM method. This test case therefore shows that by using the above LR/SG method, there exists an inherent gap (defined as the gap between the optimal dual cost and optimal feasible cost). It also shows that a tighter bound is obtained when the operation precedence constraints are not relaxed.

Case 2:   This test case is to demonstrate that the LR/BDP/ICG method outperforms the LR/SG method and the FCFS/SPT method. In this case, there are three machine types with one machine each, and four parts with a total of twelve operations. For all the parts, the due date and weight are -1 and 5 respectively, and there is no part earliness penalty. Operation processing times and required machine types are listed in table 2a, and the time horizon is 30. Testing results are summarized in tables 2b and 2c.

Since all the weights are integer, the cost of a feasible schedule should also be an integer. With the lower bound 2374.7 and feasible cost 2375 obtained by LR/BDP/ICG, the schedule obtained must thus be optimal. It can also be seen from table 2b that the LR/BDP/ICG method significantly speeds up convergence.

This case is also tested by using the FCFS/SPT method. In the FCFS/SPT method, operations are performed according to the FCFS rule. When several operations come to a machine at the same time, the SPT rule is used to determine the sequence of the these operations. The cost of the resulting schedule is 2852, which is significantly higher than the optimal feasible cost ($= 2375$). The schedule obtained by FCFS/SPT as well as the optimal schedule is presented in table 2c.

The following two cases draw data from industries, and are tested by using the LR/BDP/ICG method. The purpose is to demonstrate that the method is applicable for solving practical scheduling problems. A few performance metrics as well as the feasible costs at some iterations are evaluated to measure the schedule quality. The metrics are defined below.

Makespan: the duration of time for processing all the parts.

Maximum work-in-process inventory: the maximum number of parts in processing at a time unit.

Average work-in-process inventory: the average number of parts in processing over the makespan.

Average lead time: the average elapse time between part beginning and completion times for all parts.

Table 2a. Input data for case 2 ($\gamma = 0.5$).

Part	(Processing time/Machine type)		
	Operation 0	Operation 1	Operation 2
Part 0	4/0	3/1	2/2
Part 1	1/1	4/0	4/2
Part 2	3/2	2/1	3/0
Part 3	3/1	3/2	1/0

Table 2b. Testing results for case 2.

Method	Iteration	Dual cost	J	Duality gap	CPU (s)
LR/SG	2000	2179	2375	9.0%	4.0
LR/BDP/ICG	100	2374.7	2375	0.1%	0.1

Table 2c. Feasible schedules for case 2.

(i, j)	h	LR/BDP/ICG		FCFS/SPT	
		b_{ij}	c_{ij}	b_{ij}	c_{ij}
(0, 0)	0	0	3	0	3
(0, 1)	1	4	6	6	8
(0, 2)	2	7	8	12	13
(1, 0)	1	3	3	0	0
(1, 1)	0	4	7	4	7
(1, 2)	2	9	12	8	11
(2, 0)	2	0	2	0	2
(2, 1)	1	7	8	4	5
(2, 2)	0	9	11	9	11
(3, 0)	1	0	2	1	3
(3, 1)	2	3	5	4	6
(3, 2)	0	8	8	8	8

Case 3: This case is to show that the method selects machine types for individual operations as well as their beginning times. In this case, there are eleven machine types with a total of 16 machines and 18 parts with various due dates and weights. A part may have up to 17 operations, and the total number of operations is 159. An operation may be performed on one of up to four different machine types. With a time horizon 1086, the testing results are summarized in table 3a, and the performance metrics in table 3b.

It can be seen that both the lower bound and feasible schedule keep on improving as the number of iterations increases. It is also shown in table 3b that the schedule obtained at iteration 400 has better performance than the schedule at iteration 1.

Case 4: This case is to demonstrate the capability of the LR/BDP/ICG method for scheduling problems with large sizes. In this case, there are eight machine types with a total of 14 machines. Each machine type has one or two identical machines. A total of 82

Iteration	Dual cost	J	Duality gap	CPU (s)
1	19339	86988	350%	0.7
200	37394	40715	8.9%	128
400	37456	40170	7.2%	261

Table 3b. Performance metrics.

Iteration	1	200	400
Makespan	782	544	546
Maximum work-in-process inventory	18	16	16
Average work-in-process inventory	8.16	8.20	8.09
Average lead time	354	247	245

operation can be performed on a specific machine type. Testing results are summarized in table 4a, and the performance metrics in table 4b.

The results in tables 4a and 4b show the iterative improvement of the LR/BDP/ICG method on dual costs, feasible costs and performance metrics.

From the results of cases 3 and 4, it can be seen that significant improvement on the schedule quality and lower bound is obtained in the first 200 iterations. The improvement slows down from iteration 200 to 400. It is thus not needed to run the algorithm for a long time to get high quality schedules.

As mentioned above, the computation complexity of the BDP algorithm for solving part subproblems is $O(K \sum_{j=1}^{J_i} |H_{ij}|)$. It is observed from testing that the resolution of subproblems takes most of the CPU time in LR/BDP/ICG. The complexity of the LR/BDP/ICG algorithm is thus dominated by BDP. As shown in table 5, the computation time for 200 iterations increases almost linearly with $K \sum_{j=1}^{J_i} |H_{ij}|$.

Lagrange multipliers reflect the "price" information for using machines. In day-to-day scheduling, the multipliers associated with the previous schedule can initialize the algorithm to generate a new schedule. Since the production in a job shop may not change much from day-to-day, the computation time for finding a good schedule will

Table 4a. Testing results for case 4 ($\gamma = 0.1$).

Iteration	Dual	J	Duality gap	CPU (s)
1	84.5	46487	54914%	3
200	28837	38843	34.7%	542
400	29921	37161	24.0%	1086

Table 4b. Performance metrics of feasible schedule.

Iteration	1	200	400
Makespan	1049	1050	1047
Maximum work-in-process inventory	42	30	32
Average work-in-process inventory	22.3	15.4	16.8
Average lead time	322	219	240

Table 5. Computation time analysis for cases 3 and 4.

Case	$\sum_{j=1}^{J_i} |H_{ij}|$	K	$K \sum_{j=1}^{J_i} |H_{ij}| \ (\times 10^3)$	CPU (s)/200 iter.
3	269	1086	292	128
4	752 (269 × 2.8)	2068 (1086 × 1.9)	1555 (292 × 5.3)	542 (128 × 4.2)

decrease vastly (roughly by 2/3 according to testing experience) with initialization procedure.

5. Conclusions

In this paper, near-optimal solution methodologies for job shop scheduling are examined, and many insights are provided on a few selected methods for solving subproblems and for updating multipliers. A new algorithm is presented that combines backward dynamic programming for solving low level subproblems and interleaved conjugate gradient method for solving the high level problem. The new method significantly improves algorithm convergence and solution quality. Numerical testing for practical data sets shows that the LR/BDP/ICG method can generate high quality schedules in a timely fashion, and it is practical for job shop scheduling in industries.

References

Adams J, Balas E, Zawack D 1988 The shifting bottleneck procedure for job shop scheduling. *Manage. Sci.* 34: 391–401

Baker K 1974 *Introduction to sequencing and scheduling* (New York: Wiley)

Bertsekas D P 1995 *Nonlinear programming* (Belmont, MA: Athena Scientific)

Blackstone J H, Phillips D T, Hogg G L 1982 A state-of-the-art survey of dispatching rules for manufacturing job shop operations. *Int. J. Product. Res.* 20: 27–45

Chen H, Chu C, Proth J M 1995 A more efficient Lagrangian relaxation approach to job-shop scheduling problems. *Proc. IEEE Int. Conf. on Robotics and Automation*, pp 496–501

Czerwinski C, Luh P B 1994 Scheduling parts with bills of materials using an improved Lagrangian relaxation technique. *IEEE Trans. Robotics Autom.* 10: 99–111

Fisher M L 1973 Optimal solution of scheduling problems using Lagrange multipliers. Part I. *Oper. Res.* 21: 1114–1127

Kaskavelis C A, Caramanis M C 1995 Efficient Lagrangian relaxation algorithms for real-life-size job-shop scheduling problems. Working Paper, Department of Manufacturing Engineering, Boston University; also personal communications

Kuziak A 1990 *Intelligent manufacturing systems* (Englewood Cliffs, NJ: Prentice-Hall)

Luh P B, Hoitomt D J 1993 Scheduling of manufacturing systems using the Lagrangian relaxation technique. *IEEE Trans. Autom. Control.* 38: 1066–1079

Luh P B, Gou L, Odahara T, Tsuji M, Yoneda K, Hasegawa T, Kyoya Y 1995 Job shop scheduling with group-dependent setups, finite buffers, and long time horizon. *Proceedings of the 34th Conference on Decision and Control*, New Orleans, LA, pp 4184–4189

Luh P B, Chen D, Thakur L S 1997a Modelling uncertainty in job shop scheduling. *Proc. of the First International Conference on Operations and Quantitative Management*, Jaipur, India, pp 490–497

Luh P B, Wang J H, Wang J L, Tomastik R N 1997b Near optimal scheduling of manufacturing systems with presence of batch machines and setup requirements. *Ann. CIRP* 46: (to appear)

Pinedo M 1995 *Scheduling – Theory, algorithms and systems* (Englewood Cliffs, NJ: Prentice Hall)

Tomastik R N, Luh P B 1993 The facet ascending algorithm for integer programming problems. *Proc. IEEE Conf. on Decision and Control*, San Antonio, Texas, pp 2880–2884

Tomastik R N, Luh P B 1996 A reduced-complexity bundle method for maximizing concave non-smooth functions. *Proc. of the 31st Conference on Decision and Control*, Kobe, Japan, pp 2114–2119

Ventura J A, Weng M X 1995 Minimizing single-machine completion time variance. *Manage. Sci.* 41: 1448–1455

On-line maintenance of optimal machine schedules

AMRIL AMAN[1], ANANTARAM BALAKRISHNAN[2] and
VIJAY CHANDRU[3]

[1] FMIPA IPB, Jalan Raya Padjadjaran, Bogor, Indonesia
[2] Smeal College of Business Administration, Penn State University, University
Park, PA 16803, USA
[3] Department of Computer Science and Automation, Indian Institute of Science,
Bangalore 560 012, India
e-mail: anantb@psu.edu; chandru@csa.iisc.ernet.in

Abstract.   Effective and efficient scheduling in a dynamically changing environment is important for real-time control of manufacturing, computer, and telecommunication systems. This paper illustrates the algorithmic and analytical issues associated with developing efficient and effective methods to update schedules on-line. We consider the problem of dynamically scheduling precedence-constrained jobs on a single processor to minimize the maximum completion time penalty. We first develop an efficient technique to reoptimize a rolling schedule when new jobs arrive. The effectiveness of reoptimizing the current schedule as a long-term on-line strategy is measured by bounding its performance relative to oracles that have perfect information about future job arrivals.

Keywords.   Scheduling; design and analysis of algorithms; heuristics.

1.   Introduction

Planning and scheduling dynamic systems with random job arrivals, failures, and preemption are very challenging tasks. Typically, since future events cannot be forecast with enough detail and accuracy, planners often use on-line scheduling strategies. Consider, for instance, a production system with random job arrivals. On-line methods apply when detailed information regarding a job's processing requirement is revealed only at its release time. Thus, each release time represents an epoch at which the existing schedule is revised to reflect the new information. One on-line scheduling strategy consists of reoptimizing the current "rolling" schedule at each job arrival epoch using a deterministic scheduling algorithm that only uses information about the current system and workload status. We refer to this scheduling strategy as *on-line reoptimization*. This strategy of reacting to changes in system status (job arrival or completion, processor failure etc.) by reoptimizing and updating the current schedule raises two issues.

First, given an optimal schedule of n tasks, can we devise an *efficient* method to revise this schedule, say, when a new task enters the system? Intuitively, since the existing n-job schedule contains useful information, exploiting this information to adjust the schedule is likely to be more efficient compared to reconstructing the optimal $(n + 1)$-job schedule from scratch. We refer to the latter method as a *zero-base* algorithm, while a method that exploits current schedule information is an *updating* algorithm. Computer scientists have emphasized this issue of relative efficiency of updating methods in the context of certain geometric and graph problems by developing specialized data structures and updating algorithms (see, for example, Spira & Pan 1975, Chin & Houck 1978, Even & Shiloach 1981, Overmars & van Leeuwen 1981, Frederickson & Srinivas 1984, and Frederickson 1985). In contrast, the research on deterministic resource scheduling (see, for example, Graham *et al* 1979) focuses primarily on *zero-base* algorithms. In this paper, we illustrate the algorithmic issues in dynamic reoptimization by developing an efficient *updating* method for one class of single-machine scheduling problems.

Efficient schedule updating methods are especially important for real-time planning and control. Consider, for instance, the following "bidding" scheme for assigning tasks in a distributed processing system (see, for example, Ramamritham & Stankovic 1984, Zhao & Ramamritham 1985, Malone *et al* 1988). Jobs with varying processing requirements and due dates arrive randomly at different processor locations. Each processor maintains and updates its own local schedule. When a new job enters the system (or when a processor fails), the source node (or a central coordinator) queries the other processors to determine their expected completion time before deciding where to dispatch the job. To formulate its response, each target processor must adjust its current schedule to accomodate the new job and determine its tentative completion time. Subsequently, when the job is awarded to a processor, the selected processor must again update its schedule. Given the possibly large volume of job announcements and reassignments, devising efficient updating algorithms to accomodate new jobs is clearly critical for this type of real-time control mechanism.

In addition to updating efficiency, we are also interested in the *effectiveness* of the on-line reoptimization strategy. In particular, what is the relative performance (i.e., closeness to optimality) of schedules obtained through on-line reoptimization compared to an "optimal" off-line decision procedure that has perfect information about the future? Recently, computer scientists have developed a standardized approach to study this performance characteristic of on-line methods. The approach seeks a worst-case measure called *competitiveness* to evaluate solution effectiveness. We illustrate this mode of analysing effectiveness using our single machine scheduling example.

Studying efficiency and effectiveness issues for on-line reoptimization required a judicious choice of the scheduling context. In particular, both the structure and performance of on-line updating methods depend strongly on the scheduling objective. Consider, for instance, the problem of scheduling a single machine to minimize maximum tardiness for unrelated jobs with identical release times. The earliest due-date rule finds the optimal schedule for this problem (Jackson 1955). Given an earliest due-date (EDD) schedule for n jobs, the updating problem consists of constructing a new EDD schedule (by adjusting the current schedule) to accomodate a new job. If n denotes the number of currently scheduled

jobs, re-sorting the $(n + 1)$ jobs to construct the new EDD schedule requires $O(n \log n)$ operations. However, updating can be performed much more efficiently if we use a heap structure (see, for example, Tarjan 1983 and Aho *et al* 1974) to store the current schedule. Updating the heap when a new job arrives merely involves inserting the new item and rebalancing the heap, which requires $O(\log n)$ effort. For other scheduling objectives, the updating method is not so obvious, and the n-fold computational improvement may not be possible. And, of course, for NP-hard scheduling problems (e.g., minimizing sum of completion times for jobs with arbitrary release dates) the updating problem is not likely to be polynomially solvable either.

In this paper we develop and analyse the worst-case performance of an on-line updating method for a single machine-scheduling problem with precedence-constrained jobs, where the objective consists of minimizing the maximum completion time penalty over all jobs. In the classification scheme proposed by Graham *et al* (1979), we consider the $1/\text{prec}/f_{\max}$ problem. Lawler (1973) proposed an $O(n^2)$ zero-base algorithm to construct an optimal n-job schedule for this problem. Subsequently, Baker *et al* (1983) generalized this algorithm to the case where jobs have arbitrary but known release dates, and preemption is permitted. For the $1/\text{prec}/f_{\max}$ problem, we focus on a special class of penalty functions that satisfy a consistency property defined in § 2. Several penalty functions such as linear completion time and tardiness penalties satisfy this property. For this class of scheduling problems, § 3 first describes a new zero-base algorithm called the *Forward* algorithm (unlike Lawler's algorithm, this method schedules jobs from front to back) with $O(m + n \log n)$ worst-case time-complexity, where m denotes the number of arcs in the precedence graph. Subsequently, we develop an *updating version* of the Forward algorithm that uses information about the current n-job optimal schedule to optimally add a new job. If the new job has $n'(\leq n)$ ancestors, and the precedence subgraph induced by these ancestors has $m'(\leq m)$ arcs, the computational complexity of the Forward updating algorithm is $O(m' + n')$. Results of computer simulations reported in § 4 confirm that, in practice, the Forward updating procedure requires significantly lower computational time than applying the zero-base algorithm to construct the $(n + 1)$-job optimal schedule. Section 5 analyses the competitiveness of on-line reoptimization for selected penalty structures. We show that the method is 2-competitive (i.e., its worst-case performance ratio relative to the optimal, perfect information schedule is bounded above by a factor of 2) for a delivery-time version of the scheduling problem (without preemption), and when the penalty function is subadditive (with preemption).

This paper makes several specific contributions for the $1/\text{prec}/f_{\max}$ problem with consistent penalty functions. In particular, we: (i) propose a new Forward algorithm; (ii) develop an updating (on-line) version of the Forward algorithm; (iii) empirically demonstrate the computational benefits of using updating algorithms (instead of zero-base algorithms) to perform schedule adjustments; and, (iv) analyse the competitiveness of on-line reoptimization for some special cases. However, our broader purpose is to use the $1/\text{prec}/f_{\max}$ problem as an example to motivate the need for further work in the general area of efficient and effective schedule updating methods, and to illustrate the issues that arise in

2.   Problem description and notation

The $1/\text{prec}/f_{\max}$ problem consists of scheduling n jobs on a single machine, subject to precedence constraints on the jobs. Let p_j denote the processing time required for job j. The job precedence constraints are specified via a directed, acyclic *precedence graph G* whose nodes correspond to the jobs; the graph contains a directed arc (i, j) from node i to node j if job i is an immediate predecessor of job j. For convenience, assume that jobs are indexed from 1 to n, with $i < j$ if job i precedes job j. Let m denote the number of arcs in the precedence graph. We assume that the precedence graph is stored as a linked list requiring $O(m)$ storage, with pointers from every job to each of its immediate predecessors. Let B_j denote the set of all *immediate predecessors* of job j. Job i is said to be an *ancestor* of job j if the precedence graph contains a directed path from i to j. Let $A_j \supseteq B_j$ denote the set of all ancestors of job j. Each job j carries a non-decreasing *penalty function* $f_j(t_j)$ that depends on its completion time t_j. The scheduling objective is to *minimize* $f_{\max} = \text{Max}\{f_j(t_j) : j = 1, 2, \ldots n\}$.

Lawler (1973) developed the following $O(n^2)$ zero-base algorithm to solve the $1/\text{prec}/f_{\max}$ problem. The method iteratively builds an optimal sequence by scheduling jobs in reverse order, i.e., it first identifies the job to be processed last (i.e., in position n of the schedule), then the job in position $(n - 1)$, and so on. At stage k, let Q_k denote the set of k currently unscheduled jobs, and let T_k denote the cumulative processing time for all jobs in Q_k, i.e., $T_k = \Sigma\{p_j : j \in Q_k\}$. Also, let R_k be the subset of jobs in Q_k whose successors, if any, have all been already scheduled; we refer to jobs in R_k as the set of *eligible* jobs at stage k. During stage k, the algorithm assigns to position k the eligible job $j^* \in R_k$ with minimum penalty at time T_k, i.e., $f_{j^*}(T_k) = \text{Min}\{f_j(T_k) : j \in R_k\}$. The procedure terminates at the end of stage 1. Since each step requires $O(n)$ operations to identify the eligible job with minimum penalty, the overall complexity of the algorithm is $O(n^2)$.

While Lawler's algorithm applies to arbitrary penalty functions, we will focus on a special class of penalty functions that satisfy the following *consistency* condition: A set
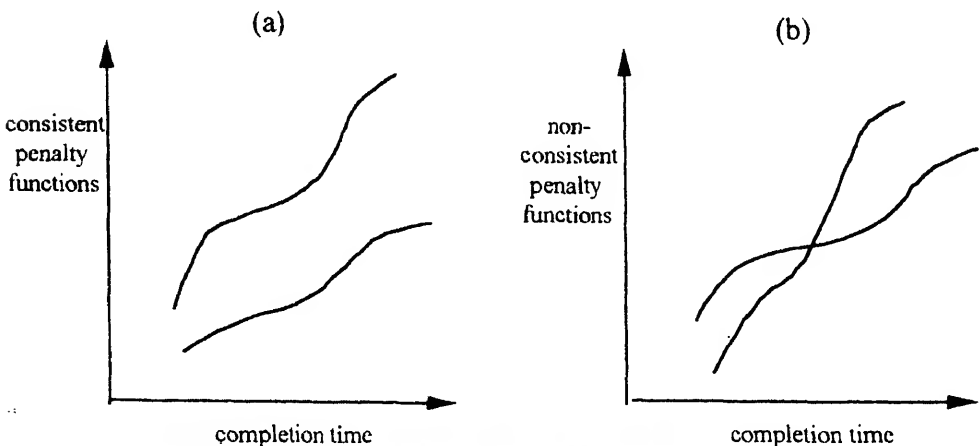
Figure 1. Consistent (a) and non-consistent (b) penalty functions
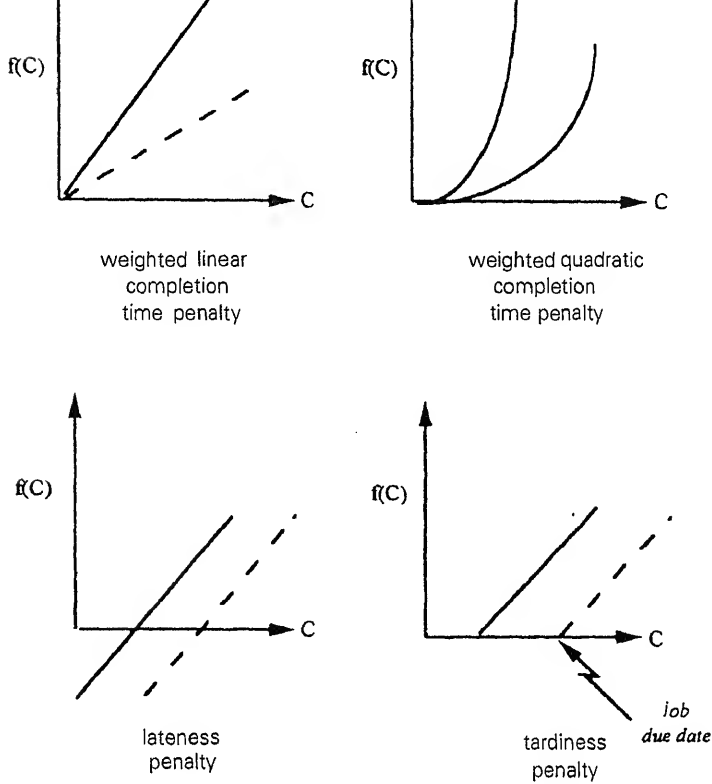
Figure 2. Special consistent penalty functions.

of functions $f_1(.)$, $f_2(.)$, ..., $f_n(.)$ is said to be *consistent* if, for every pair of indices $i, j \in \{1, 2, \ldots, n\}$, either $f_i(t) \leq f_j(t)$ or $f_i(t) > f_j(t)$ *for all values of completion time t*.

As figure 1a shows, consistent functions do not intersect; Figure 1b shows two penalty functions that are not consistent. Several natural penalty functions satisfy the consistency property. Examples, shown in figure 2, include (i) the *weighted* (linear and quadratic) *completion time* criteria $f_j(t) = w_j t$ or $f_j(t) = w_j t^2$ (ii) the *lateness penalty* $f_j(t) = t - d_j$, where d_j is the due date for job j, and (iii) the *tardiness penalty* $f_j(t) = \max\{0, t - d_j\}$. Jobs with consistent penalty functions have the same relative ranking (say, increasing order of penalties) for all completion time values. Hence, we will sometimes omit the completion time argument, and denote as $f_i > f_j$ the fact that job i has a higher penalty than job j (for all completion time values).

For convenience, we will assume that the penalty functions for the different jobs are distinct, i.e., either $f_i > f_j$ or $f_i < f_j$. Thus, the job with the maximum penalty will always be unique. Our Forward algorithm requires only the relative order of jobs with respect to the penalty functions rather than the exact penalty values for different completion times. Hence, ranking the jobs in order of penalties is sufficient.

This section first describes a new zero-base procedure called the *Forward* algorithm to find the optimal n-job schedule for the 1/prec/f_{max} problem with consistent penalty functions. Unlike Lawler's algorithm, the new method schedules jobs from front to back (i.e., it assigns jobs to earlier positions first). We prove the correctness of this algorithm, and demonstrate how it facilitates updating the schedule when a new job enters the system.

3.1 The Forward zero-base algorithm

The Forward algorithm is motivated by the following intuitive argument. Recall that, for consistent penalty functions, the relative ordering of jobs (in terms of their penalties) does not vary with time. Hence, if jobs are not constrained by precedence restrictions, we can minimize f_{max} by scheduling the jobs in decreasing order of penalty. However, this decreasing-penalty order may violate some precedence constraints. To satisfy the precedence constraints, consider the following 'natural' scheme to selectively (and parsimoniously) advance jobs: Start with the decreasing-penalty order as the candidate sequence; examine jobs from front to back in this sequence, and ensure precedence feasibility for each job j by advancing (i.e., scheduling immediately before job j) every ancestor that is currently scheduled after job j. This procedure effectively attempts to deviate as little as possible from the decreasing-penalty order by advancing only the essential low-penalty jobs that must precede the high-penalty jobs. As we demonstrate next, this principle forms the basis for the Forward algorithm, and gives the optimal n-job schedule.

To describe and prove the validity of the Forward algorithm, we use some additional notation. For any subset of jobs S, let $B_j(S)$ be the set of all immediate predecessors of job j belonging to subset S, i.e., $B_j(S) = B_j \cap S$. Similarly, $A_j(S) = A_j \cap S$ denotes the set of ancestors of job j in subset S. The Forward algorithm relies on the following result. Recall that we have indexed jobs such that $i < j$ if job i precedes job j.

PROPOSITION 1
For any subset of jobs S, let j^* be the job in this subset with the maximum penalty. Then, subset S has an optimal schedule, denoted as $\Pi(S)$, that assigns job j^* to position $\{|A_{j^*}(S)| + 1\}$, and all its ancestors to the first $|A_{j^*}(S)|$ positions in increasing order of job indices (where $|A|$ denotes the number of elements of the set A).

Proof. The first part of the proposition states that the subset S must have an optimal schedule that processes the maximum-penalty job j^* as soon as possible, i.e., this schedule first processes all ancestors of job j^*, followed immediately by j^*. We prove this result using an interchange argument. Consider an alternative optimal schedule Π'' that does not satisfy this property. Let job j^* be scheduled in position $k > |A_{j^*}(S)| + 1$, and let job $j' \notin A_{j^*}(S)$ be a non-ancestor that is scheduled closest to, but before, job j^*. Let k' denote the position of job j' in schedule Π'', $k' < k$. By our choice of k', all jobs in positions $(k' + 1)$ to $(k - 1)$ must be ancestors of j^*. Also, job j' is not an ancestor for any of these jobs; otherwise, j' would be j^*'s ancestor as well. Finally, since job j^* has the maximum penalty in the set S, $f_{j^*}(t) > f_{j'}(t)$, where t is the current completion time

of job j^*. Consider now the new schedule obtained by postponing job j' to position k, and advancing all jobs in positions $k' + 1$ to k by one position. By our previous observations, the new schedule must be feasible; furthermore, since job j^* has a higher penalty than job j', the new schedule does not increase the maximum penalty of the schedule. By repeating this process until all non-ancestors of job j^* are postponed beyond j^*, we get an optimal schedule that satisfies the condition of the proposition.

Now, job j^* has the maximum penalty among all jobs in S. Thus, the penalty incurred for j^* must exceed the penalty for each of its ancestors (since these are completed earlier and have lower penalty functions), regardless of their relative order in positions 1 to $|A_{j^*}(S)|$. To be feasible, however, the assignment of these ancestors must satisfy the precedence constraints. Since jobs are numbered in order of their precedence, scheduling the ancestors in increasing index order gives a feasible schedule. $\square$

Proposition 1 suggests the following iterative scheduling procedure: first, identify the job j^* with maximum penalty, schedule it in position $(|A_{j^*}(S)| + 1)$, and assign all its predecessors to positions 1 through $|A_{j^*}(S)|$. Let $S' \subseteq S$ denote the remaining set of jobs (which are not ancestors of j^*). In the optimal schedule $\Pi(S)$, these remaining jobs must be scheduled optimally in positions $(|A_{j^*}(S)| + 2)$ to $|S|$. In effect, we can consider a new scheduling problem for the subset of jobs S', and apply proposition 1 to this new subset, and so on. Our method implements this iterative procedure. We formally describe the algorithm next. In this description, r is the *iteration counter*, l_r is the pointer to the last position in the schedule that is filled in the rth iteration, and S_r is the set of remaining unscheduled jobs at the beginning of iteration r.

The Forward zero-base algorithm

Step 0: Initialization

Set $r \leftarrow 1$; iteration counter
$S_r \leftarrow \{1, 2, \ldots, n\}$; set of unscheduled jobs at iteration r
$l_{r-1} = 0$. last position scheduled in previous iteration

Step 1: Iterative step

(a) Find the job $j^*(r)$ with maximum penalty in set S_r;
$A_{j^*(r)}(S_r) :=$ Set of all ancestors of $j^*(r)$ in S_r.
(b) Set $l_r \leftarrow l_{r-1} + |A_{j^*(r)}(S_r)| + 1$;
Assign $j^*(r)$ to position l_r
(c) Assign jobs of the set $A_{j^*(r)}(S_r)$ to positions $(l_{r-1} + 1)$ through $l_r - 1$ in increasing order of job indices.
(d) Set $S_{r+1} \leftarrow S_r - A_{j^*(r)}(S_r) - \{j^*(r)\}$
(e) If S_{r+1} is empty, Stop. The current schedule is optimal;
Else, set $r \leftarrow r + 1$, and return to step 1(a).

Observe that the iterative step is performed at the most n times. We refer to the job $j^*(r)$ with the maximum penalty at the rth step as the rth *bottleneck job*. As r increases, the corresponding bottleneck jobs have successively lower penalties.
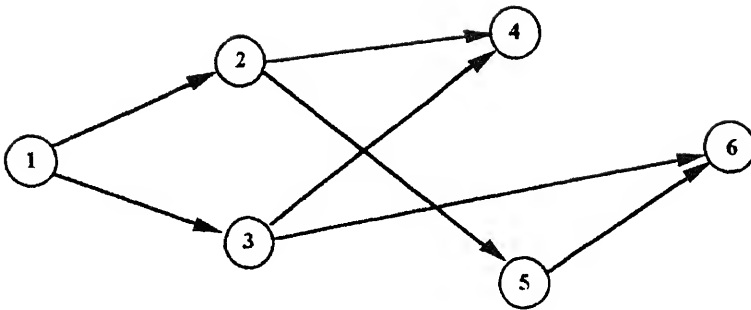
Ranking of jobs in order
of decreasing penalties :   5, 1, 6, 3, 2, 4

n = *number of jobs* = 6
m = *number of arcs in precedence graph* = 7

Figure 3.   Example for Forward zero-base algorithm.

3.1a   *Example*:   To illustrate the Forward algorithm, consider the precedence graph and the relative ordering of 6 jobs (in decreasing order of penalties) shown in figure 3.

Initially, all jobs are unscheduled, and the job with the largest penalty is job 5 (i.e., $j^*(1) = 5$). This job has two unscheduled ancestors, jobs 1 and 2, i.e., $A_5(S_1) = \{1, 2\}$. The first iteration schedules the ancestors in positions 1 and 2, and schedules job 5 in position $l_1 = 3$. At the second iteration, job 6 has the largest penalty among all remaining jobs. Its unscheduled ancestor, job 3, is assigned to position 4, while job 6 is scheduled in position 5. In the final iteration, the only remaining job (job 4) is scheduled in the last position. Table 1 summarizes these computations.

3.1b   *Data structures and computational complexity*:   To perform the Forward algorithm's computations efficiently, we maintain a special data structure, and make some minor algorithmic changes. Our implementation first sorts the jobs in decreasing order of penalties prior to initiating the main algorithm. This sorting operation requires $O(n \log n)$ effort, and will facilitate the process of identifying the bottleneck job at each iteration of the main procedure. Also, our implementation does not determine the exact positions for the unscheduled ancestors in the set $A_{j^*(r)}(S_r)$ (i.e., it does not perform step 1(c)) immediately after each iteration. Instead, we reindex these jobs temporarily, and determine the actual

Table 1.   Forward zero-base algorithm iterations for example.

Iteration k	Unscheduled jobs S_k	Bottleneck job $j^*(k)$	Unscheduled ancestors $A_{j^*(k)}(S_k)$
1	1, 2, 3, 4, 5, 6	5	1, 2
2	3, 4, 6	6	3
3	4	4	—

Optimal schedule: 1–2–5–3–6–4

final schedule at the end of the main algorithm by performing an overall sorting operation. Initially, all jobs have a temporary index of 0. At iteration r, we assign the temporary index $(nr + j)$ to each job $j \in A_{j^*(r)}(S_r)$ that is scheduled during that iteration, and job $j^*(r)$ is assigned the index $(nr + j^*(r))$. Thus, all jobs that must be scheduled in the rth iteration (between positions $(l_{r-1} + 1)$ and $(l_r - 1)$) have temporary indices in the range $(nr + 1)$ to $n(r + 1)$. After the main algorithm terminates, we sort the jobs in increasing order of their temporary indices ($O(n \log n)$ effort) to obtain the final optimal schedule. Observe that the largest possible value of a temporary index is $n(n + 1)$. Also, at intermediate iterations, all the jobs that have not yet been scheduled are easy to identify since they have temporary indices of 0.

Let us now analyse the computational complexity of the Forward algorithm. First, identifying the successive bottleneck jobs involves sequentially scanning the sorted list of jobs, which requires $O(n)$ total effort. (At step r, the rth bottleneck job $j^*(r)$ is the first job following $j^*(r - 1)$ in the sorted list with a temporary index of 0.) Now, consider the effort required to identify the unscheduled ancestors of job $j^*(r)$ (in step 1(a)). Starting with job $j^*(r)$, we trace back all unscheduled ancestors using the pointers to the immediate predecessors in the linked list representation of the precedence graph. If we encounter a previously scheduled job, we need not explore its ancestors since these ancestors must all be scheduled previously. Thus, the total effort required to identify the members of the set $A_{j^*(r)}(S_r)$ is $O(m)$ over all iterations (since we examine each edge in the precedence graph exactly once). Combined with the initial and final sorting operations, we get an overall complexity of $O(m + n \log n)$. In general, the number of arcs m in the precedence graph is $O(n^2)$; hence, the Forward algorithm is no better than Lawler's original algorithm in the worst-case. However, for problems with sparse precedence graphs, we expect the Forward algorithm to perform better.

The Forward algorithm also extends to the more general $1/\text{prec}, r_j, \text{pmtn}/f_{\max}$ problem where jobs have different release times r_j that are known in advance, and preemption is permitted. Appendix 1 describes this extension. Later (in § 5), we use the schedule generated by this enhanced method as the benchmark to evaluate the effectiveness of on-line reoptimization when job release times are not known in advance.

3.1c *Adapting Lawler's algorithm for consistent penalty functions*: Note that when the penalty functions are consistent, we can also adapt Lawler's original $O(n^2)$ algorithm for $1/\text{prec}/f_{\max}$ to run in $O(m + n \log n)$ time. Recall that, at each stage k, for $k = n$, $n - 1, \ldots, 1$, Lawler's algorithm selects the eligible job $j \in R_k$ with the smallest penalty at the current completion time T_k. For general penalty functions, the order of eligible jobs (arranged in increasing order of penalty values at T_k) might change from stage to stage. However, with consistent penalty functions, the order is invariant. To exploit this property we use a heap structure to store the currently eligible jobs in sorted order (increasing penalties) at each stage. At stage k, we – (i) schedule the job that is currently at the root of the heap, (ii) delete this job from the heap, and (iii) insert in the heap all its immediate predecessors that just became eligible. Inserting and removing each job from the heap entails $O(n \log n)$ total effort; and, checking the eligibility of jobs at each stage requires $O(m)$ effort (since each arc of the precedence graph must be examined once).

Hence, the overall complexity of the heap implementation of Lawler's static algorithm is $O(m + n \log n)$, which is the same as the computational complexity of our Forward algorithm. However, as we show next, the Forward algorithm is more amenable to updating existing schedules.

3.2 The Forward updating algorithm

For the *updating* problem, we are given an optimal n-job schedule, and a new job, indexed as $(n + 1)$, with prespecified immediate predecessors B_{n+1}, arrives. We initially assume that job $(n + 1)$ does not have any successors in the current set of n jobs; later, we indicate how to apply the updating method when the new job also has successors among currently scheduled jobs. Let $\Pi = \{j_1, j_2, j_3, \ldots, j_n\}$ denote the current optimal n-job schedule where j_k denotes the index of the job that is scheduled in the kth position. The updating problem consists of constructing a new optimal schedule $\Pi' = \{j'_1, j'_2, \ldots, j'_n, j'_{n+1}\}$ that includes the new job $(n + 1)$.

The updating procedure uses information on the bottleneck jobs corresponding to the current schedule. As we mentioned in § 3.1, the successive bottleneck jobs must have successively lower penalties; hence, the current sequence Π already lists the bottleneck jobs in order of decreasing penalties. Consider now the position for job $(n + 1)$ in the $(n + 1)$-job optimal schedule Π', assuming we applied the Forward zero-base algorithm. Since job $(n+1)$ does not precede any current jobs, its position in the schedule is determined solely by its penalty function relative to the current bottleneck jobs. In particular, suppose $f_{j*(r-1)} < f_{n+1} < f_{j*(r)}$, i.e., the new job's penalty lies between the penalties for the $(r - 1)$th and rth bottleneck jobs. Since the current sequence schedules existing bottleneck jobs in decreasing penalty order, we can identify the index r in linear time. And, the updating procedure must merely insert job $(n + 1)$ and all its previously unscheduled ancestors (i.e., ancestors that are not scheduled in positions 1 to l_{r-1}) immediately after the $(r - 1)$th bottleneck job $j^*(r - 1)$.

Assuming that we are initially given only the immediate predecessors B_{n+1} of job $(n+1)$, finding the members of $A_{n+1}(S_r)$ (the set of unscheduled ancestors for job $(n+1)$) requires $O(m')$ operations, where m' is the number of edges in the precedence subgraph induced by job $(n + 1)$ and its ancestors. We must then assign these ancestors to consecutive positions, starting with position $(l_{r-1} + 1)$. Observe that the current schedule satisfies the precedence constraints among all ancestors of job $(n + 1)$. Hence, the jobs in $A_{n+1}(S_r)$ need not be re-sorted to satisfy precedence constraints; instead, we get a feasible schedule by merely scheduling these ancestors in the order in which they occur in the current schedule. Adjusting the current schedule in this manner requires at most $O(n')$ effort, where n' is the number of ancestors of job $(n + 1)$. Thus, the overall complexity of the Forward *updating* procedure is $O(m' + n')$ compared with $O(m + n \log n)$ for the Forward *zero-base* algorithm. Finally, note that the updating procedure can easily accommodate new jobs that must precede existing jobs. Let j_s be the immediate successor of job $(n + 1)$ that is scheduled earliest in the current schedule, and let l_s denote its position in the current schedule. In the new schedule, the jobs that are currently scheduled in positions l_s and beyond will retain their relative order. Therefore, we need to apply the updating procedure only to jobs that are currently scheduled in positions 1 to $l_s - 1$.
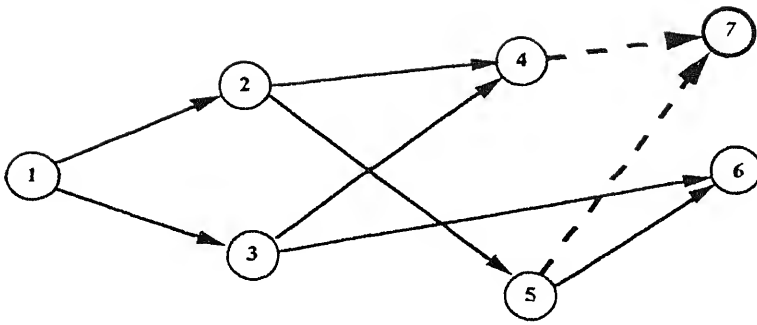
Ranking of jobs in order
of decreasing penalties : 5, 1, 7, 6, 3, 2, 4

Updated optimal schedule: 1 - 2 - 5 - 3 - 4 - 7 - 6

Figure 4. Example for Forward updating algorithm.

3.2a *Example*: For the example shown in figure 3, figure 4 illustrates the updating
calculations when a new job (job 7) enters the system. This job has two immediate prede-
cessors, jobs 4 and 5, and its penalty value lies between those of jobs 1 and 6. Job 7 must,
therefore, be scheduled between the two consecutive bottleneck jobs 5 and 6. Job 7 has two
ancestors, jobs 3 and 4, that are not scheduled prior to job 5. Hence, we assign these two jobs
to positions 4 and 5, respectively (preserving their relative order in the current schedule);
job 7 occupies position 6, followed by job 6. Figure 4 shows the updated schedule.

4. Computational results

Section 3 showed that the updating algorithm has better worst-case complexity than the
zero-base algorithm. To verify this computational superiority in practice, we compared the
computation times using the Forward zero-base algorithm and the updating procedure for
an extensive set of random test problems ranging in size from 100 jobs to 400 jobs. We first
describe the random problem generating procedure before presenting the computational
results.

4.1 *Random problem generation*

Our random problem generator requires two user-specified parameters: the number of
jobs (n), and the density δ of the precedence graph $(0 < \delta \leq 1)$. Initially, we attempted
to generate precedence graphs with random topologies by independently selecting arcs
(i, j), for any pair of nodes i and j, with probability δ. We discovered, however, that the
resulting precedence graphs contained many redundant arcs. For example, if the graph
contains arcs (i, j), (j, k), and (i, k), then arc (i, k) can be deleted since this precedence
order (i.e., i preceding k) is implied by the other two arcs. Consequently, the reduced
graph (with redundant arcs eliminated) was often much sparser than the desired density

values. To overcome this problem and to avoid checking for redundancies, we decided to use *layered* precedence graphs that contain only arcs between successive layers; hence, none of the arcs are redundant.

To generate a random layered graph containing n nodes and with density parameter δ, the problem generator

- randomly selects the number of layers (L) in the graph $(0 < L \leq n)$;
- equally divides the number of nodes among the layers; and,
- for each pair of nodes i and j in successive layers, selects arc (i, j) with probability δ.

For the updating problem, the random generator assigns the new job $(n + 1)$ to a new layer, and connects node $(n + 1)$ to nodes in the previous layer with probability δ.

We implemented the zero-base and updating versions of the Forward algorithm in PASCAL on a Sun 4/390 workstation. For our computational tests, we considered four network sizes, with number of nodes $n = 100, 200, 300$, and 400, and five values of the density parameter $\delta = 0.10, 0.25, 0.50, 0.75$, and 0.90. For each combination (n, δ), we generated 100 random problem instances. Table 2 summarizes the mean (over 100 random instances) and standard deviation of CPU times (to add the $(n + 1)$th job to the current schedule) for the zero-base and updating versions of the Forward algorithm for all the (n, δ) combinations.

As table 2 shows, the updating algorithm is faster than the zero-base version by a factor ranging from 2 to 5. Thus, for scheduling contexts that require numerous, frequent updates, the magnitude of computational savings using the updating method can be substantial. For the 100-job problems, the CPU time for individual problem instances varies widely as indicated by the large standard deviation (relative to the mean). As the problem size increases, the CPU time for updating relative to zero-base scheduling appears to increase. The density parameter does not seem to have a significant or consistent effect on this ratio.

Table 2. Computation times for updating versus zero-base algorithm.

Arc density (d)	Number of jobs $(n) =$							
	100		200		300		400	
	Zero-base	Update	Zero-base	Update	Zero-base	Update	Zero-base	Update
0.90	24.9[†]	5.0	76.7	25.3	141.6	58.6	222.0	102.4
	(10.7)[§]	(7.7)	(13.0)	(8.9)	(19.6)	(9.0)	(27.8)	(12.6)
0.75	27.6	6.7	76.3	23.4	138.7	58.3	219.0	103.4
	(7.8)	(8.3)	(12.1)	(9.5)	(19.6)	(9.7)	(24.3)	(8.9)
0.50	24.6	5.7	73.9	29.9	132.7	56.0	212.1	100.0
	(8.3)	(8.0)	(14.3)	(11.2)	(17.5)	(8.7)	(23.8)	(7.5)
0.25	23.7	7.7	66.3	26.4	124.7	55.7	200.0	102.3
	(9.5)	(8.4)	(10.4)	(8.9)	(11.7)	(9.4)	(22.6)	(10.1)
0.10	24.4	5.7	66.3	25.3	123.9	58.1	186.0	104.0
	(9.0)	(8.1)	(11.8)	(8.2)	(12.1)	(10.2)	(15.5)	(13.8)

[†] Average CPU time (in milliseconds on a SUN 4/90) over 100 random problem instances;
[§] Standard deviation of CPU times

5. Competitiveness of on-line reoptimization

Having demonstrated the relative *efficiency* of using tailored updating methods instead of zero-base algorithms to accomodate new jobs, we now examine the effectiveness of on-line reoptimization as a heuristic strategy to schedule dynamic systems. One approach to evaluate this effectiveness is to assume a tractable stochastic model for job arrivals, processing times, and precedence relationships, and analyse the expected performance of on-line reoptimization (or other dispatch rules) in this framework. However, this mode of analysis is often sensitive to the choice of the stochastic model governing the occurrence of random events.

Recently several researchers in theoretical computer science (e.g., Borodin *et al* 1987, Manasse *et al* 1988, Chung *et al* 1989) have developed an alternate approach to study on-line effectiveness using the notion of *competitiveness*. The approach involves characterizing the worst-case performance of the on-line method compared to an optimal off-line procedure that has perfect information about the future. In particular, for problems with a minimization objective, an on-line algorithm A is said to be *c-competitive* if the inequality,

$$C_A \leq c\,C_0 + \alpha,$$

holds for any instance of the on-line problem. Here, C_A denotes the "cost" incurred by the on-line algorithm A, and C_0 is the cost for the optimal off-line solution with clairvoyance. Thus, competitiveness is a useful measure for performance analysis of incremental algorithms. Researchers have started applying this measure to scheduling problems only recently; Shmoys *et al* (1991) address competitiveness issues related to on-line scheduling of parallel machines to minimize makespan.

This section demonstrates the underlying principles and techniques of competitiveness analysis applied to our single machine scheduling problem. Developing bounds on the relative difference between the on-line and off-line optimal objective values for *general* penalty functions $f_j()$ is difficult since we cannot exploit any special properties of the solutions. We, therefore, need to separately study various specialisations of $f_j()$. We consider two types of penalty functions – subadditive functions and lateness. For subadditive penalty functions, we show that on-line reoptimization is 2-competitive when we permit preemptions, and $(\rho + 2)$-competitive for non-preemptive scheduling, where ρ is the aspect ratio (defined later). We then prove 2-competitiveness of on-line reoptimization for the lateness penalty case (delivery-time version).

Before describing and proving the competitiveness results, let us clarify the context and the mechanics of on-line reoptimization. Jobs arrive randomly at various *release times r_j*. Assume that jobs are indexed in the order in which they arrive. We study effectiveness for both preemptive and non-preemptive scheduling problems. First, consider the case when preemptions are permitted, i.e., at each arrival epoch r_j, the job that is currently in process, say, job u can be interrupted and resumed later without any additional setup or reprocessing effort. In this case, applying the updating method involves: (i) determining the rank, say, r^* (in decreasing penalty order) of the new job j relative to all the currently available jobs (including the current in-process job u), and (ii) inserting job j and its unscheduled ancestors immediately after the $(r^* - 1)$th bottleneck job in the current schedule. Notice that the current job u is preempted only if job j has a higher penalty than all other available

jobs. In the non-preemptive case, job u must necessarily be completed first, and only the remaining jobs can be rescheduled. Hence, job j is ranked only relative to these remaining jobs.

This on-line updating algorithm is a heuristic method that does not guarantee long-run optimality of the schedules. The benchmark for comparing the performance of the on-line method is an optimal off-line schedule that has prior knowledge (at time 0) about the exact arrival times r_j for all jobs j. In Graham *et al*'s (1979) nomenclature, the off-line schedule is the optimal solution to either the $1/\text{prec}, r_j, \text{pmtn}/f_{\max}$ or $1/\text{prec}, r_j/f_{\max}$ problem depending on whether or not preemption is permitted. Note that the $1/\text{prec}/f_{\max}$ problem (with preemption) is polynomially solvable using, say, the enhanced Forward algorithm described in appendix A, while the $1/\text{prec}, r_j/f_{\max}$ problem (without preemption) is known to be NP-hard. Indeed, the non-preemptive problem remains NP-hard even if we restrict the penalty $f_{\max}$ to maximum lateness $L_{\max}$, and relax the precedence constraints (i.e., for the $1/r_j/L_{\max}$ problem).

5.1 *Subadditive penalty functions*

A penalty function $f_j()$ is said to be *subadditive* if it satisfies the condition

$$f_j(t_1 + t_2) \leq f_j(t_1) + f_j(t_2), \quad \text{for all } t_1, t_2 \, {}^3 0.$$

Interesting special cases of subadditive functions include *concave* penalty functions, and linear completion time penalties (i.e., $f_j(t) = \beta_j t$). We denote the maximum subadditive penalty as $f_{\max}^{SA}$. This section studies the competitiveness of on-line reoptimization, with and without preemption, when all jobs have consistent, subadditive penalty functions. When preemptions are permitted, we show that the on-line reoptimization strategy is 2-competitive, i.e., the objective value of the on-line schedule is at most twice the optimal off-line value. When preemption is prohibited, the worst-case ratio increases to $(\rho + 2)$, where ρ is a specified ratio of job processing times.

5.1a *Scheduling with preemptions*: We now show that, for any k-job problem, the preemptive schedule obtained using on-line reoptimization (without anticipating future job arrivals) has a worst-case performance ratio of 2 relative to the optimal off-line schedule for the $1/r_j$, prec, $\text{pmtn}/f_{\max}^{SA}$ problem constructed by the enhanced Forward algorithm (appendix A).

Theorem 1. *For the $1/\text{prec}, r_j, \text{pmtn}/f_{\max}^{SA}$ problem, on-line reoptimization is 2-competitive.*

Proof. Let Π be the preemptive schedule obtained using on-line reoptimization for a k-job problem. Let ϕ_k be the (maximum) completion time penalty for this schedule, and let j' denote the *critical* job, i.e., $\phi_k = f_{j'}(t_{j'}) = \max\{f_j(t_j) : 1 \leq j \leq k\}$, where t_j is the completion time for job j in the schedule Π. First, we note several characteristics of the schedule Π. Since j' is the critical job, all jobs following job j' in schedule Π must have lower penalty functions (otherwise, a job with higher penalty function that is scheduled later than j' would be the critical job). Consider now the interval of time $[r_{j'}, t_{j'}]$ between the arrival of job j' and its completion. Let J' denote the set of all jobs j that are completed

in this interval and *having equal or higher penalty functions* (J' also includes job j). Let $A(J')$ denote the set of ancestors $j \in J'$ of all jobs in the set J'. Clearly, every job that the on-line schedule completes in the interval $[r_{j'}, t_{j'}]$ must either be a member of the set J' or an ancestor of some job $j \in J'$. Also, the completion time $t_{j'}$ for job j' in schedule Π has the following upper bound:

$$t_{j'} \leq r_{j'} + \sum_{j \in J'} p_j + \sum_{j \in A(J')} p_j. \tag{1}$$

Now consider the optimal off-line schedule Π^* which uses prior information on job release times. Let ϕ_k^* be the (maximum) completion time penalty for this schedule, and let T_j be the completion time for job j in Π^*. Among all the jobs belonging to the set J', let j'' be the job that is scheduled last in Π^*. Observe that

$$\phi_k^* \geq f_{j''}(T_{j''}) \geq f_{j'}(T_{j''}), \tag{2}$$

$$T_{j''} \geq \sum_{j \in J'} p_j + \sum_{j \in A(J')} p_j, \text{ and} \tag{3}$$

$$T_{j''} \geq r_{j'}. \tag{4}$$

Inequality (2) follows from the definition of ϕ_k^*, and because job $j'' \in J'$ has an equal or higher penalty function than job j'. Inequalities (3) and (4) hold because both j' and j'' belong to the set J', and job j'' is completed last in Π^*. From (1), (3), and (4), we have

$$t_{j'} \leq T_{j''} + T_{j''} = 2T_{j''}. \tag{5}$$

Inequalities (2) and (5) imply that

$$\begin{aligned}
\phi_k &= f_{j'}(t_{j'}) \\
&\leq f_{j'}(2T_{j''}), \quad \text{from (5),} \\
&\leq 2f_{j'}(T_{j''}), \quad \text{from subadditivity, and} \\
&\leq 2\phi_k^*, \quad \text{from (2).}
\end{aligned} \tag{6}$$

Hence, the on-line reoptimization strategy is 2-competitive. $\square$

Claim. The worst-case bound of 2 (proved in theorem 1) is tight.

The following example justifies this claim. Consider a 3-job problem instance with linear completion time penalties $f_j(t) = \beta_j t$. Job 1 is unrelated, while job 2 precedes job 3. Jobs 1 and 2 arrive at time 0, each requiring 10 units of processing time, and job 3 arrives at time 10 with a very small processing time. The jobs have the following ordering in terms of penalty functions: $f_3 > f_1 > f_2$. In our notation, the following parameters describe this problem instance: $r_1 = 0$, $r_2 = 0$, and $r_3 = 10$; $p_1 = 10$, $p_2 = 10$, and $p_3 = \varepsilon$; $B_3 = \{2\}$; and $\beta_1 = 1$, $\beta_2 = 0$, and $\beta_3 = 100$.

At time 0, both jobs 1 and 2 are available, but job 1 has higher penalty. Hence, the on-line method processes it first. The next epoch is at time 10, when job 3 arrives (and job 1 completes). At this time, the on-line method starts job 2 (to satisfy job 3's precedence constraint), and finally completes job 3 at time $(20 + \varepsilon)$. Job 3 is the critical job, with a

method anticipates job 3's higher penalty and delayed arrival at $t = 10$. Hence, the optimal off-line sequence is 2–3–1. Job 3, completed at $(10 + e)$, is again the critical job, with a penalty of $100^*(10 + \varepsilon)$. Thus, the ratio of on-line to off-line penalty values is $(20 + \varepsilon)/(10 + \varepsilon)$ which approaches 2 as ε approaches 0. Hence, the worst-case ratio of 2 is tight. Next, we show that, for the non-preemptive case, the worst-case ratio is higher.

5.1b *Scheduling without preemptions*:   As we noted earlier, finding the optimal off-line schedule with no preemption is computationally intractable. Hence, unlike the preemptive case, we do not have a convenient characterization of the optimal off-line schedule. To evaluate the competitiveness of on-line reoptimization for the non-preemptive case, we use the following strategy. We know that the optimal off-line preemptive schedule has a lower penalty than the optimal off-line non-preemptive schedule. Hence, if we can derive a worst-case ratio for the *on-line, non-preemptive schedule* with respect to the *optimal, off-line preemptive schedule*, this ratio should also hold for the off-line non-preemptive solution. Let ρ denote the maximum, over all job pairs i and j, of the ratio of processing time for job j to the sum of processing times for job i and all its ancestors, i.e.,

$$\rho = \max \left\{ p_j \middle/ \left\{ p_i + \sum_{l \in A(i)} p_l \right\} : 1 \le i, j \le k, i \ne j \right\}.$$

We refer to ρ as the *Aspect ratio*. Note that the denominator in the above expression is a lower bound on the earliest possible completion time of job i. Indeed, our competitiveness result applies even if we replace the denominator of ρ with a tighter lower bound (on job i's completion time) involving, say, the release times of job i and its ancestors.

Theorem 2. *For the $1/prec, r_j/f_{\max}^{SA}$ problem, the on-line reoptimization method is* $\max\{3, (\rho + 2)\}$*-competitive.*

Proof. This proof is very similar to the proof for theorem 1. Let Π_{np} be the on-line, non-preemptive schedule for a k-job problem. Let ϕ_k^{np} be the (maximum) penalty of this schedule, defined by the critical job j'. Consider the interval of time $[r_{j'}, t_{j'}]$ between the arrival of job j' and its completion in schedule Π_{np}. Let u be the job that is currently in process in Π_{np} when job j' arrives, and let J' be the set of all jobs with equal or higher penalties than j' that are completed in the time interval $[r_{j'}, t_{j'}]$. Except for the in-process job u, all other jobs that are completed in this interval either belong to J' or are ancestors of one or more jobs in J'. Let $A(J')$ be the set of all ancestors $j \notin J'$ for jobs in J'. As before, let Π^* be the optimal, off-line *preemptive* schedule; job $j'' \in J'$ is scheduled last in Π^* among all jobs of J'. T_j denotes the completion time for job j in Π^*, and ϕ_k^* is the schedule's penalty value. The job completion times in the on-line and off-line schedules must satisfy the following inequalities:

$$t_{j'} \le r_{j'} + p_u + \sum_{j \in A(J')} p_j + \sum_{j \in J'} p_j; \tag{7}$$

$$T_{j''} \ge \max \left\{ r_{j'} \sum_{j \in A(J')} p_j + \sum_{j \in J'} p_j \right\}; \text{ and} \tag{8}$$

$$\phi_k^* \geq f_{j''}(T_{j''})$$

$$\geq f_{j'}\left(\max\left\{r_{j'}\sum_{j\in A(J')}p_j + \sum_{j\in J'}p_j\right\}\right). \tag{9}$$

These inequalities imply that:

$$\phi_k^{np} = f_{j'}(t_{j'})$$

$$\leq f_{j'}(r_{j'}) + f_{j'}(\rho p_{j'}) + f_{j'}\left(\sum_{j\in A(J')}p_j + \sum_{j\in J'}p_j\right)$$

using subadditivity and (7)

$$\leq \phi_k^* + \max\{\phi_k^*, \rho\phi_k^*\} + \phi_k^* \text{ using subadditivity and (9)}$$

$$= \max\{3, (\rho+2)\}\phi_k^*.$$

When $\rho \leq 1$, the competitive ratio is 3 and is otherwise no larger than $(\rho+2)$. Thus, the worst-case ratio for the on-line, non-preemptive schedule is at most $\max\{3, (\rho+2)\}$ relative to the optimal off-line, preemptive schedule. Hence, the on-line reoptimization strategy is at least $\max\{3, (\rho+2)\}$-competitive for the $1/r_j$, prec, pmtn/$f_{\max}^{SA}$ problem. $\qquad\square$

Claim. Assuming $\rho \geq 1$, the worst-case bound of $\max\{3, (\rho+2)\}$ for non-preemptive schedules is tight.

To prove this claim, consider the following augmented version of the previous worst-case problem instance (described after theorem 1). In addition to the 3 jobs in that example, we have a fourth job that arrives at time $r_4 = 0$, with penalty coefficient $\beta_4 = 1$, and a processing time of $p_4 = 20$. Also, job 3 arrives at $r_3 = 10 + \delta$, for some small $\delta > 0$. Note that the aspect ratio ρ for this problem instance is $p_4/p_1 = 2$. Consider, first, the schedule obtained using on-line reoptimization. Job 1 (or job 4) is scheduled first and completes at time 10. Since job 3 is not yet available, the method schedules job 4, followed by job 2 and finally job 3. Job 3 completes at time $(40 + \varepsilon)$; it is the critical job with a penalty of $100^*(40 + \varepsilon)$. Contrast this on-line schedule with the following optimal, *non-preemptive* schedule: Job 2 starts at time 0 and completes at time 10; the processor is idle from time 10 to time $(10 + \delta)$; Job 3 starts at time $(10 + \delta)$, and completes at time $(10 + \delta + \varepsilon)$, followed by jobs 1 and 4. Again, job 3 is the critical job, with a completion time penalty value of $100^*(10 + \delta + \varepsilon)$. Thus, the ratio of the on-line penalty and the optimal, off-line (non-preemptive) penalty approaches $(\rho + 2) = 4$ as δ and ε tend to zero.

5.2 The Lateness penalty function

We now consider the *Lateness* objective $L_{\max}$, i.e., the penalty for job j is $f_j(t_j) = t_j - d_j$, where t_j and d_j are, respectively, the completion time and due date for job j, and $L_{\max} = \text{Max}\{f_j(t_j) : j = 1, 2, \ldots, n\}$. Our discussions focus on non-preemptive, precedence-constrained scheduling for this problem. The best off-line approximation algorithm for the $1/r_j/L_{\max}$ problem was developed by Hall & Shmoys (1992). They show

that, for the delivery-time formulation of this model (described later), a 4/3 approximation algorithm is possible using an enhanced version of Jackson's earliest due date rule. Although Jackson's rule can be applied on-line, the approximation techniques used by Hall and Shmoys sacrifice the on-line characteristic to achieve the tighter bounds.

To study on-line competitiveness, we cannot work directly with the maximum lateness objective L_{max} since some problem instances may have zero or negative optimal off-line L_{max} values. (If the off-line L_{max} is zero, the worst-case competitiveness becomes unbounded for any on-line algorithm that is even slightly suboptimal.) Instead of redefining competitiveness, we transform the L_{max} problem to the following equivalent delivery time version (Potts 1980) which has a positive optimal value for all problem instances.

5.2a *The delivery time formulation*:

Let $\{d_j\}$ denote the job due dates, and let K be a value greater than the largest due date. Now define the *tail* q_j of job j as

$$q_j = K - d_j.$$

We interpret the tail q_j as the time to deliver the job j after it is completed. Thus, the delivery time of job j is $t_j + q_j$, where t_j is the completion time of job j. The objective of the scheduling problem now consists of minimizing the maximum delivery time over all jobs. The optimal schedule for this delivery time version also minimizes maximum lateness. Observe that the penalty function $f(t_j) = t_j + K - d_j$ implied by the delivery time objective function is consistent (according to our definition in § 2), with jobs that are due earlier having higher penalty functions. Thus, at each job arrival epoch, on-line reoptimization for the delivery time problem (without preemptions) involves adjusting the current schedule to process the available job with the earliest due date as soon as possible (subject to completing the current in-process job and all unfinished ancestors of the EDD job).

5.2b *2-Competitiveness of on-line reoptimization*:

For any given k-job instance of the non-preemptive delivery time problem, let Π be the schedule obtained using on-line reoptimization. Let DT_j be the delivery time of job j in this schedule; ϕ_k is the maximum delivery time value over all jobs j. Denote the (maximum) delivery time of the optimal, off-line schedule Π^* as ϕ_k^*.

Theorem 3. *On-line reoptimization is 2-competitive for the non-preemptive delivery time problem, i.e., $\phi_k \leq 2\phi_k^*$ for all k.*

Proof. We prove this result by induction. Assume that jobs are indexed in the order in which they arrive, and first consider $k = 2$. Since delivery time for a job equals the sum of its start time, processing time, and tail, the optimal off-line value ϕ_k^* must be greater than or equal to the maximum processing time over the k jobs. In the two-job case, if both jobs arrive simultaneously, then the on-line updating method also constructs the optimal schedule. When the release times are different, say, job 1 arrives before job 2, the on-line method schedules job 1 before job 2. Suppose the optimal off-line solution consists of processing job 2 before job 1. Relative to this optimal schedule, the on-line schedule

2's penalty in 11' is a lower bound on the optimal off-line objective function value ϕ_2^*. Hence,

$$\phi_2 \leq \phi_2^* + p_1 \leq \phi_2^* \max\{p_1, p_2\}$$
$$\leq 2\phi_2^*.$$

Thus, the on-line method is 2-competitive for 2 jobs. Now suppose the method is 2-competitive for all $k' \, 2(k-1)$ jobs. We show that it must also be 2-competitive for k jobs.

Let job $j' \leq k$ be the critical job in the on-line schedule, i.e., $\phi_k = DT_{j'} = s_{j'} + p_{j'} + q_{j'}$, where s_j is the start time for job j in the on-line schedule.

Case 1. $s_{j'} < s_k$. In this case, job j' starts earlier than the new job k but also has the highest delivery time. Hence, job j' must have a higher penalty function (i.e., earlier due date) than job k. Since the Forward algorithm leaves all higher penalty jobs unaffected when job k arrives, job j' must start at $s_{j'}$ even in the $(k-1)$ job on-line schedule. Hence, $\phi_{k-1} \geq s_{j'} + p_{j'} + q_{j'} = \phi_k$. But, $\phi_{k-1}^* \leq \phi_k^*$, and $\phi_{k-1} \leq 2\phi_{k-1}^*$ by the induction hypothesis. Hence, $\phi_k \leq 2\phi_k^*$.

Case 2. $s_{j'} \geq s_k$. In this case,

$$\phi_k^* \geq r_{j'} + p_{j'} + q_{j'}$$
$$= (s_{j'} + p_{j'} + q_{j'}) - (s_{j'} - r_{j'})$$
$$= \phi_k - (s_{j'} - r_{j'}).$$

Note however that $s_j - r_j \leq \sum_{i=1}^{j} p_i$ for all jobs j.

Otherwise, the interval of time between the arrival and start of job j contains some idle time which is impossible using the on-line method (since job j is waiting in the queue). Therefore, $\phi_k^* \geq \phi_k - \sum_{i=1}^{k} p_i$, which implies that $\phi_k \leq 2\phi_k^*$.

Thus, on-line reoptimization is 2-competitive. $\qquad\square$

As before, we can show that the worst-case ratio of 2 (proved in theorem 3) for on-line reoptimization is tight. Indeed, for the non-preemptive delivery time problem, the following example (Kise & Uno 1978; Potts 1980) shows that *any* on-line scheduling algorithm that does not introduce forced idleness (i.e., does not keep the machine idle when the job queue is not empty) must have a worst-case ratio of *at least 2*. Consider a problem instance with two jobs that are released respectively at $r_1 = 0$ and $r_2 = 1$, having processing times $p_1 = (P-1)$ and $p_2 = 1$, and due dates $d_1 = P$ and $d_2 = 1$ (hence, the tails are $q_1 = 0$ and $q_2 = (P-1)$). With no precedence constraints, the best strategy consists of keeping the machine idle for the first time period, and scheduling job 2 before job 1. The maximum delivery time for this solution is $(P+1)$. On the other hand, any on-line method that does not anticipate future job arrivals or introduce forced idleness will schedule job 1 at time 0 (since it is the only available job). Since jobs cannot be preempted, job 2 begins processing

only at time $(P-1)$, and its delivery time is $(2P-1)$. As P becomes arbitrarily large, the ratio of on-line to optimal off-line delivery time approaches 2. And, on-line reoptimization achieves this lowest possible worst-case ratio.

In retrospect, the 2-competitiveness of on-line reoptimization is not surprising in view of the worst-case bound of 2 for Schrage's heuristic (Potts 1980). For the $1/r_j/DT_{\max}$ delivery time problem (without preemption or precedence constraints), Schrage's heuristic consists of applying Jackson's earliest due date rule on-line (with a longest processing time tie-breaking rule), i.e., whenever a job completes, the method dispatches the currently available job with the earliest due date. Note that, without precedence constraints, on-line reoptimization also chooses this "current" EDD job sequence. Potts (1980) used a characterization of Schrage's heuristic schedule (in terms of an *interference* job) to prove that the method has a worst-case bound of 2. When jobs have precedence constraints, we can transform the delivery time problem $1/r_j$, prec$/DT_{\max}$ to an equivalent unconstrained version by revising the job release times and tails as follows: If job i must precede job j, set $r_j \leftarrow \max\{r_i, r_j\}$ and $q_i \leftarrow \max\{q_i, q_j + p_j\}$. Potts' result implies that Schrage's heuristic applied to this transformed problem is 2-competitive. Note that, when a new job arrives, updating the tails (to account for its precedence constraints) involves examining every currently available ancestor of the new job. In contrast, our on-line updating algorithm (appendix A) is more efficient since it first locates the new job relative to the current bottleneck jobs, and only examines ancestors that are scheduled later.

6. Conclusion

In this paper we have developed a new Forward algorithm and an updating version for one class of scheduling problems. The updating procedure reduces the computational effort to accomodate a new job into an existing schedule by using information from the current schedule. In contrast, applying a zero-base algorithm to reschedule all the jobs from scratch would entail significantly higher computational effort, as illustrated by our computational results of § 4. Section 5 gives partial results characterizing the effectiveness of using deterministic reoptimization for on-line scheduling.

Our broader purpose in this paper is to demonstrate the scope, and efficiency and effectiveness issues in developing on-line updating algorithms for dynamic scheduling problems. In spite of their practical importance in contexts such as real-time control of distributed processors, updating algorithms have not been adequately studied in the scheduling literature. For illustrative purposes, we studied a single-machine scheduling problem that can be solved efficiently. Exploring similar updating methods for other scheduling objectives and contexts is an important research direction that merits further investigation.

Appendix A. Forward algorithm for the $1/\text{prec}, r_j, \text{pmtn}/f_{\max}$ problem with consistent penalties

The Forward algorithm described in § 3 assumes that all jobs are simultaneously available at time 0. This appendix describes an extension to handle arbitrary, but known, job release times; job preemption is permitted.

First, we review the notation. We are given n jobs, and a precedence graph $G : (N, A)$ containing m arcs. B_j and A_j represent, respectively, the set of immediate predecessors and ancestors of job j. Each job has a "consistent" non-decreasing penalty function $f_j(t)$, i.e., either $f_j(t) < f_i(t)$ or $f_j(t) > f_i(t)$ for all completion times t, and $f_j(t) \geq f_j(t')$ if $t > t'$. Let p_j and r_j denote, respectively, the processing time and release time for job j.

Without loss of generality, we assume that all ancestors of any job j arrive at or before job j; otherwise, we can set $r_j = \text{Max}\{r_i + p_i : i \in B_j\}$. Also, for convenience, assume that jobs are indexed in order of release dates; consequently, $i < j$ if job i precedes job j. We require a preemptive schedule that minimizes $f_{\max} = \text{Max}\{f_j(t_j) : j = 1, 2, \ldots, n\}$ while satisfying the precedence constraints and release times, where t_j is the completion time for job j in the chosen schedule.

Scheduling principle　As before, the Forward algorithm identifies successive bottleneck jobs, and schedules each bottleneck job as early as possible. As before, we first sort all jobs in decreasing order of penalties. The method starts with an empty schedule, and progressively assigns jobs to appropriate "free" time intervals in the current schedule. At iteration r, the method has scheduled the first $(r-1)$ bottleneck jobs and all their ancestors. Let I^r denote the set of available free time intervals in the current schedule at the start of iteration r. The following steps are performed at iteration r:

Step 1. Identify the rth bottleneck job $j^*(r)$, i.e, job $j^*(r)$ has the largest penalty among all the currently unscheduled jobs;

Step 2. Consider each unscheduled ancestor j of job $j^*(r)$ in increasing index order: allocate to job j the first available p_j time units *after the release time* r_j. Update the available free intervals.

Step 3. Allocate to job $j^*(r)$ the first available $p_{j^*(r)}$ time units after the release time $r_{j^*(r)}$.

The algorithm iteratively repeats this process – finding the next bottleneck job, and scheduling this job and all its ancestors as early as possible in increasing index order. Note that as we identify and schedule more bottleneck jobs, the schedule becomes fragmented, i.e., it has "holes" due to delayed releases for certain jobs. These holes are filled whenever possible in subsequent steps; filling the holes might introduce preemptions, i.e., the total processing time of a job may be distributed over several intervals.

Let Π be the final schedule constructed by the Forward algorithm, and denote the set of bottleneck jobs as J_B.

Lemma.　*In the final schedule constructed by the Forward algorithm,*

$$f_{\max} = \text{Max}\{f_j(t_j) : j \in J_B\}.$$

Proof of correctness of the Forward algorithm. Suppose the schedule Π constructed by the Forward algorithm is not optimal. Let j^* be the *critical* job that determines the penalty of this schedule, i.e., $f_{\max} = f_{j^*}(t_{j^*})$, and suppose j^* is the r^*th bottleneck job, i.e., $j^* = j^*(r^*)$. Let Π' be an optimal schedule; let $f'_{\max}$ and t'_j denote, respectively, the maximum penalty and the completion time for each job j in this schedule.

By the hypothesis, $f'_{\max} < f_{\max} = f_{j^*}(t_{j^*})$. Since penalty functions are non-decreasing, this inequality implies that job j^* must be scheduled earlier in Π', i.e., $t'_{j^*} < t_{j^*}$. Since the Forward algorithm schedules all bottleneck jobs as early as possible in order of decreasing penalties, t'_{j^*} can be less than t_{j^*} only if the schedule Π' completes some previous bottleneck job (i.e., a bottleneck job with a higher penalty than job j^* *and* scheduled before t_{j^*} in the Forward schedule Π) on or after t_{j^*}. Let $j^*(r)$ for some $r < r^*$ denote this bottleneck job. Since $t'_{j^*(r)} > t_{j^*}$, we must have $f'_{\max} \geq f_{j^*(r)}(t_{j^*})$. However, since the job $j^*(r)$ has a higher penalty than j^*, $f_{j^*(r)}(t_{j^*}) > f_{j^*}(t_{j^*}) = f_{\max}$, contradicting the optimality of schedule Π'. □

References

Aho A V, Hopcroft J E, Ullman J D 1974 *The design and analysis of computer algorithms* (Reading, MA: Addison-Wesley)

Baker K R, Lawler E L, Lenstra J K, Rinnooy Kan A H G 1983 Preemptive scheduling of a single machine to minimize maximum cost subject to release dates and precedence constraints. *Oper. Res.* 31: 381–386

Borodin A, Linial N, Saks M 1987 An optimal on-line algorithm for metrical task systems. *Proc. of 19th ACM Symposium on Theory of Computing*, pp 373–382

Chin F, Houck D 1978 Algorithms for updating minimum spanning trees. *J. Comput. Syst. Sci.* 16: 333–344

Chung F R K, Graham R L, Saks M E 1989 A dynamic location problem for graphs. *Combinatorica* 9: 111–131

Even S, Shiloach Y 1981 An on-line edge deletion problem. *J. Assoc. Comput. Mach.* 28: 1–4

Graham R L, Lawler E L, Lenstra J K, Rinnooy Kan A H G 1979 Optimization and approximation in deterministic sequencing and scheduling: A survey. *Ann. Discrete Math.* 5: 287–326.

Frederickson G N 1985 Data structures for on-line updating of minimum spanning trees with applications. *SIAM J. Comput.* 14: 781–798

Frederickson G N, Srinivas M A 1984 On-line updating of degree-constrained minimum spanning trees. *Proceedings of the 22nd Allerton Conference on Communication, Control, and Computing*, October (New York: IEEE Press)

Hall L A, Shmoys D 1992 Jackson's rule: Making a good heuristic better. *Math. Oper. Res.* 17: 22–35

Jackson J R 1955 Scheduling a production line to minimize maximum tardiness. Research Report 43, Management Science Research Project, University of California, Los Angeles

Kise H, Uno M 1978 One-machine scheduling problems with earliest start and due time constraints. *Mem. Kyoto Tech. Univ. Sci. Technol.* 27: 25–34

Lawler E L 1973 Optimal sequencing of a single machine subject to precedence constraints. *Manage. Sci.* 19: 544–546

Lawler E L, Lenstra J K, Rinnooy Kan A H G 1982 Recent developments in deterministic sequencing and scheduling: A survey. In *Deterministic and stochastic scheduling* (eds) M A H Dempster, J K Lenstra, A H G Rinnooy Kan (Dordrecht: Riedel)

Malone T W, Fikes R E, Grant K R, Howard M T 1988 Enterprise: A market-like task scheduler for distributed computing environments. In *The ecology of computation* (ed) B A Huberman (Amsterdam: Elsevier Science) pp 177–205

Manasse M S, McGeoch L A, Sleator D D 1988 Competitive algorithms for on-line problems. *Proc. 20th ACM Symposium on Theory of Computing* (New York: ACM Press) pp 322–333

Overmars M H, van Leeuwen J 1981 Maintenance of configurations in the plane. *J. Comput. Syst. Sci.* 23: 166–204

Potts C N 1980 Analysis of a heuristic for one machine sequencing with release dates and delivery times. *Oper. Res.* 28: 1436–1441

Ramamritham K, Stankovic J A 1984 Dynamic task scheduling in distributed hard real-time systems. *IEEE Software* 1: 96–107

Sahni S, Cho Y 1979 Nearly on line scheduling of a uniform processor system with release times. *SIAM J. Comput.* 8: 275–285

Shmoys D, Wein J, Williamson D P 1991 On-line scheduling of parallel machines, preprint

Spira P M, Pan A 1975 On finding and updating spanning trees and shortest paths. *SIAM J. Comput.* 4: 215–225

Tarjan R E 1983 *Data structures and network algorithms* (Philadelphia, PA: Soc. Ind. Appl. Math.)

Zhao W, Ramamritham K 1985 Distributed scheduling using bidding and focused addressing. *Proceedings of the Symposium on Real-time Systems* (New York: IEEE Press) pp 103–111

Advances in discrete material handling system design

SRINIVASAN RAJAGOPALAN and SUNDERESH S HERAGU

Decision Sciences and Engineering Systems Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
e-mail: [rajags,herags]@rpi.edu

Abstract. The paper presents an outline of the research done in the area of facility layout and discrete material handling system design. The objective of this paper is to observe the trend in the literature in the area of manufacturing system design and predict the direction of future research in this area. The paper attempts to link the areas of facility layout design and material flow network design. A methodology for solving the integrated design problem is presented. An algorithm which aids in solving the combined pick-up/drop-off point location and material handling flowpath problem is outlined.

Keywords. Facility layout; automated guided vehicles; material handling system; integrated system design.

1. Introduction

The key design aspects involved in a manufacturing company are product design and process design. Product and process design involve designing the product while taking into consideration the processes to be used to manufacture that product. Process design includes equipment selection, designing the machine layout and the material handling system. Two key aspects that influence the efficiency of a manufacturing setup are the facility layout and material handling system design and operation. This paper analyzes the trend in the literature pertaining to facility layout and discrete material handling systems and projects the direction of future research in this area.

2. Facility layout design

Depending on how the machines are laid out, a manufacturing system can be broadly classified as:

(a) Traditional manufacturing system;

(b) Cellular manufacturing system.

In a traditional manufacturing system, the machines are arranged within a given space to minimize the total material travel. The cellular manufacturing system is based on the

concept that all machines that perform operations on parts with similar processing require-ments can be classified and grouped into cells (Heragu 1994). A cellular manufacturing layout tries to minimize material handling costs between cells and also between the indi-vidual machines in a cell. The different methods available for solving the layout problem are the quadratic assignment problem (QAP) and its variations, graph theoretic approach, and meta-heuristics based on simulated annealing, genetic algorithms and tabu search. For a more detailed survey of the techniques available for modelling and solving the facil-ity layout problem, the reader is referred to Kusiak & Heragu (1987), Heragu & Kusiak (1991), Heragu (1992, 1997) & Hassan (1994).

The earliest formulation for the facility layout problem involved setting it up as a QAP. The objective function in this case was a quadratic function which calculates the total cost of transferring material between machines. The constraints for this model specify that each site is assigned to only one machine and that each machine can occupy only one site. The formulation includes binary variables which indicate whether or not a machine is assigned to a given site. The implicit assumptions in this model are that all the cells are of equal size and shape. The QAP is known to be NP-complete (Sahni and Gonzalez 1976). In the past, attempts to solve the QAP optimally have focussed on transforming it into an equivalent linear mixed integer model. The transformed model, which has a larger number of variables and constraints than the QAP, was then solved using branch and bound and cutting plane algorithms. These algorithms were good for solving problems involving a maximum of 10 to 15 facilities.

Exchange algorithms based on the 2-opt and 3-opt methods have also been used to solve the layout problem. These algorithms are based on the principle of exchanging the positions of 2 or 3 facilities and checking the effect of this exchange on the objective function. If an improvement in the objective function is found, then the new layout is selected. The drawback in these methods is that they do not consider all the possible layouts that exist in the solution space and therefore are sub-optimal algorithms. These algorithms need an initial starting point which can be any feasible solution. The quality of the final solution depends upon the initial solution. However, these methods allow the user to generate a few good layouts from which the final layout can be selected.

The shift in paradigm and focus on cellular manufacturing systems led to a focus on an additional area of research, namely machine grouping. The design problem is divided into two parts – (a) Grouping of machines into cells, and (b) layout of the cells. In order to generate an overall optimal layout, a solution approach which simultaneously determines the grouping of machines and the layout of the cells is desirable. However, due to the complexity in modelling and solving such a combined problem, most approaches attempt to solve the two problems sequentially. For a more detailed explanation and a literature review of the cellular manufacturing system layout problem, the reader is referred to Cheng (1995).

3. Material handling system design

material handling equipment to departmental moves and developing the flowpath for the system. The types of material handling devices to be used, are mainly dependent on the size of the company, the types of operations taking place and amount of space available on the factory floor. Smaller companies prefer to use more traditional material handling devices such as forklifts or trucks, sometimes in conjunction with pick and place robots. Assembly operations and line-balancing operations tend to use conveyors because they allow the workers to perform the necessary operations on the part while it moves from one station to another. Automated storage/retrieval systems (AS/RS) are used in warehousing operations. Industries that are involved in sorting and packaging type of operations tend to use palletizers and sortation conveyors. The current trend seen in large manufacturing companies is the employment of automated guided vehicles (AGVs) as material handling devices. AGVs are preprogrammed trucks that are controlled through a central computer. The AGVs go through a predetermined route and make stops at designated machines, where the material is either picked up or dropped off by other automated material transfer devices. Because AGVs are expected to be the dominant material handling device in future manufacturing systems, this paper focusses on the use of AGVs in the material handling function.

An AGV system comprises 3 parts, namely the vehicle, the guidance system and the control system. The vehicle in itself is a driverless unit with a built-in charging system, an on-board controller, a communication unit and a frame. The guidance system consists of a guidepath and a tracking system. The guidepath could be either a physical metal wire used for tracking (active tracking) or a wireless system which detects the vehicle based on optical or wireless metal detection techniques (passive tracking). The control system consists of the controller which brings together the vehicle and the tracking system. It acts as the brain for the entire system. The control system also takes upon the task of route selection, blocking and tracking loads. An AGV system has several advantages among which are reliability, automated operation, reduced material damage during transfer, ease of change in the layout and increased efficiency of the entire manufacturing system. It also offers the user real-time control of the material handling system and on-the-spot decisions such as change in routing or assignment of AGVs to specific moves. The main disadvantage of AGVs are their high cost. They also need polished floor surfaces and removal of obstacles in the flowpath before they can be put into operation.

A control classification scheme for an AGVS is presented by Peters *et al* (1995). A cubic classification structure that indicates the complexity of the problem based on guidepath determination, vehicle capacity and vehicle addressing mechanism is presented. Details of the levels and sublevels that partition an AGV system are also presented. This paper concentrates more on the control aspects of an AGV system. Sinreich (1996) and Heragu & Rajagopalan (1996) discuss in detail the research done in the area of AGV flowpath design. Sinreich (1996) gives an overview of the design models that have been developed for discrete material flow systems. Heragu & Rajagopalan (1996) present a classification scheme which can be used as a tool in identifying the various aspects involved in developing a flowpath for an AGV based system and also discuss the merits and demerits of all the models developed so far.

The classification scheme for the AGV flowpath design problem presented in Heragu & Rajagopalan (1996) is based on:

Type of flowpath – traditional, single loop, tandem layout, loop layout and terminal AGV flowpath;

Travel direction – unidirectional, bi-directional;

Solution techniques – mathematical programming models, rule based solutions, heuristic solution techniques, simulation methods, mixed approach (i.e. combined simulation and analytical approach);

Assumptions made – layout given, pick-up/drop-off points given;

Objective function – total loaded travel time, travel distance of loaded vehicles, total travel distance, maximum travel distance, total path distance, overall cost.

The main categories for classification are the type of flowpath assumed in the problem, the type of AGV flowpath design required – i.e., the travel direction of the AGVs, the solution techniques used to solve the models developed, the assumptions made while developing these models and the objective function used in them.

4. AGV flowpath design

Maxwell & Muckstadt (1982) first identified the problem of AGV system design. They tackled the problem of identifying the number of AGVs required for a system assuming that the guidepath was already laid out. The first models developed to solve the layout problems were mathematical programming models which tried to minimize the total travel distance of loaded vehicles for a given layout. Branch and bound techniques were developed to
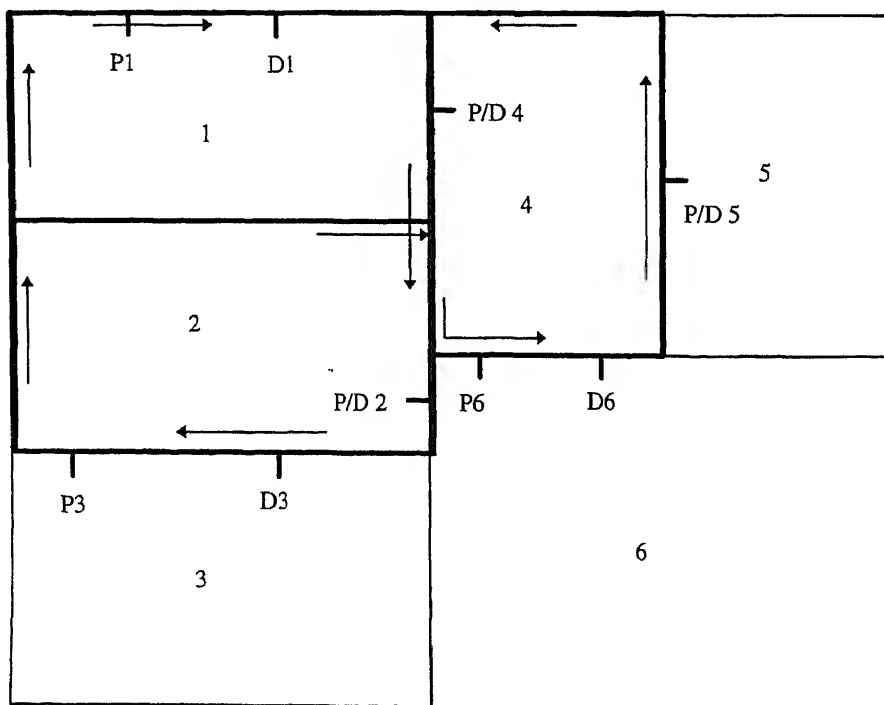
Figure 1. A schematic showing a traditional AGV flowpath design.

blems. The traditional flowpath problem was also modelled as a graph
ved. Figure 1 shows a schematic of a traditional flowpath. A graph network
ck-up/drop-off points as nodes and the flowpaths as arcs. Pick-up/drop-
:s are those points in a machine or a cell where the material handling
unloads the material required for that machine or cell. Later formulations
ze the total loaded and unloaded travel time for a given AGV system.
approaches to solving the flowpath problem also include formulating it as
lity flow problem. The solution methods to all these problems are mainly
:h are easy in solving small problems. Rule-based solutions are those which
a problem and cannot be generalized. The rules generated simplify the
consideration and reduce the size of the solution space. However most of
become impractical when solving large, real life problems (i.e. problems
20 facilities). Kouvelis *et al* (1992) developed a set of heuristics to solve
ath problem. They presented 5 different heuristic approaches to solve the
lowpath problem including a simulated annealing based approach. Usher
esented a two-phase iterative heuristic to solve the unidirectional layout
etermination of P/D points. The first phase develops a flowpath based on a
nd in the second phase the P/D points are determined based on minimizing
tween any two stations using *m* one-median problems.
velopment in AGV flowpath design came with the concept of an optimal
;ure 2). This concept developed by Tanchoco & Sinreich (1992) is an al-
: traditional flowpath design. An optimal single loop involves connecting
nts with the smallest possible single unidirectional loop. The simplicity of
iminates problems such as collision, blocking, and other routing conflicts.
ies to find the best single loop guidepath and locates the P/D points along
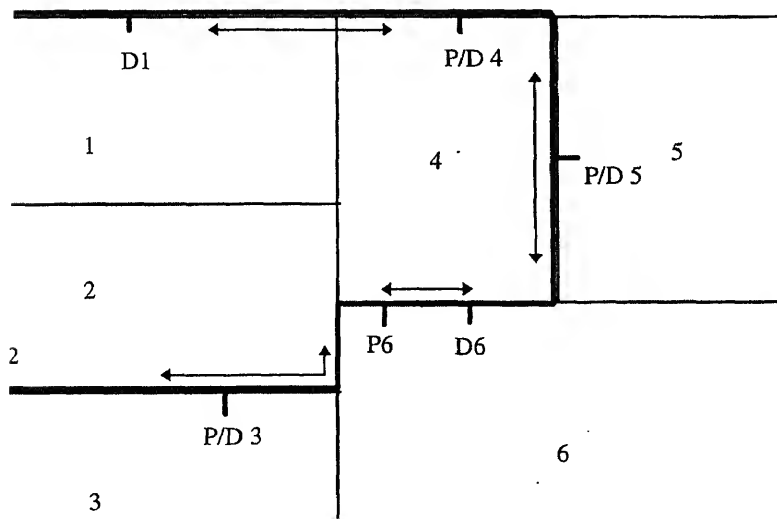
decided based on a set of rules optimizing the desired performance criteria. Sinreich & Tanchoco (1991) also developed other heuristic-based and rule-based methods to solve the P/D point location and optimal single loop layout problem. They developed a centroid projection method which considered both inter-departmental and intra-departmental flows (Sinreich & Tanchoco 1991). A genetic algorithm based method was developed by Banerjee & Zhou (1995) which designed the facility layout for a single loop-based material flow network.

The loop layout design was formulated to ease the problem of designing the facility layout and the material-handling system. The wide application of such a layout also helped in increased research in this area. The loop layout consists of machines being arranged in a loop so as to facilitate material handling. In these problems, the material-handling system is assumed to be known and in most cases is either a unidirectional AGV or a conveyor. The problem then reduces to that of a single row layout problem with certain added constraints that machines may not be placed near the edges. Most solution techniques developed to solve these problems are either heuristic-based or rule-based methods.

A concept which simplifies the material handling system design even further and makes the control aspects of an AGV somewhat easier is the tandem layout. Developing a tandem layout involves grouping machines into cells and having a single loop AGV for each cell. Material transfer stations aid in transferring material from one cell to another. Bozer & Srinivasan (1991) first discussed the concept of a tandem layout. They also developed a strategy for partitioning the shop floor into appropriate cells and developed a facility layout. The tandem layout assumes that each cell is serviced by a single unidirectional AGV. This design increases the chance of the entire production line being disrupted if any of the AGVs breaks down. Performance evaluation of a tandem layout versus the conventional AGV system has been done using petri-net models and simulation. The studies conclude that a tandem layout, under normal operating conditions, performs just as efficiently as a traditional layout and is also easier from a control perspective. Wang & Hafeez (1994) proposed a terminal-based AGV system, where the AGVs are located in a terminal and on receiving a signal, the first available AGV goes to the pickup station along a preassigned path and then proceeds to the dropoff point. On completion of the task, the AGV returns to the main terminal. Variations of the terminal based AGV system were developed by Johnson & Brandeau (1995). They used the principles of queuing network theory to evaluate their system.

Apart from the different types of layouts and flowpaths discussed above, the area of AGV flowpath design also considered the travel direction as an important parameter. Most studies assumed that the travel direction of the AGV was unidirectional, i.e. it could travel in one direction only. Egbelu & Tanchoco (1986) conducted one of the earliest studies on the merits and demerits of using bi-directional travel as opposed to unidirectional travel. The problem of conflict routing and collision avoidance is more acute in the case of bi-directional AGVs. The solution methods used to solve this problem include techniques such as column generation (Krishnamurthy et al 1993), simulation (Faraji & Batta 1994), graph theoretic approaches (Dowsland & Greaves (1994)) and expert systems (Dhiouib & Kadi (1994). The scheduling of AGVs has also been studied widely by researchers (Ulsoy & Bilge 1993; Blazewicz et al 1994). The different scheduling rules, their relationship with

the routing problem and the effect of these on the overall operation of the manufacturing system has been researched using techniques such as simulation and queuing theory.

5. Integrated design process

In the current literature, the most common assumptions made when solving the flowpath problem are that the layout is already given and that the P/D points are already determined. The papers that deal with the loop layout problem assume that the material handling system has already been designed and then solve the layout problem. The tandem layout structure simplifies the flowpath problem, but relatively more AGVs are needed for a tandem layout than for a conventional layout and therefore it becomes a somewhat expensive proposition. Also, a breakdown of one of the vehicles in the system could bring the entire production to a halt.

The current manufacturing design process involves a sequential process where the initial stage involves process selection and equipment selection. This is followed by the layout design phase. Once the layout has been decided, the P/D points are selected for this layout and then a material handling system is designed to minimize total travel time for the given layout and P/D point locations. In such a case, the system that is designed may not be optimal overall although each phase in itself might be optimal. In other words, the AGV flowpath designed is optimal for the given layout, but the combined layout and flowpath may not be globally optimal. In order to overcome this, a model which simultaneously solves the flowpath and the layout must be developed. However, this combined problem is known to be NP-complete (Banerjee & Zhou 1995) and solving such a model may not be possible.

The area of integrated design has been tackled by a few researchers. Chajjed *et al* (1992) solved the problem of identifying the P/D points and developed a free-flow material-handling, network assuming the block layout was already known. Wu & Egbelu (1994) tackled the problem of concurrently determining the layout and material-handling design by using a multistage approach. They use a loop-layout model and solve the problem of locating the cells within the loop. In this case the material-handling flowpath design problem is a trivial problem which translates into determining the P/D points and layout along a loop. Palliyil & Goetschalckx (1996) used a dual simplex-based approach to solve the problem of determining the aisles and load stations for an aisle-based material-handling system. They assume that the layout is already known. Other works which tackle the problem of integrating two aspects of a manufacturing design problem include the combined layout and P/D point location problem (Heragu 1990), flowpath design and load size determination (Egbelu 1990) and equipment selection in conjunction with the load size determination (Noble *et al* 1994).

Integrated design problems are part of the current area of research which is being widely pursued. Heragu and Rajagopalan (1996) present a framework for developing a near-optimal-solution to the manufacturing system design problem (figure 3). The framework involves a multi-stage solution procedure. The proposed strategy is a three-stage methodology that should result in an optimal or good solution depending on the size of the problem
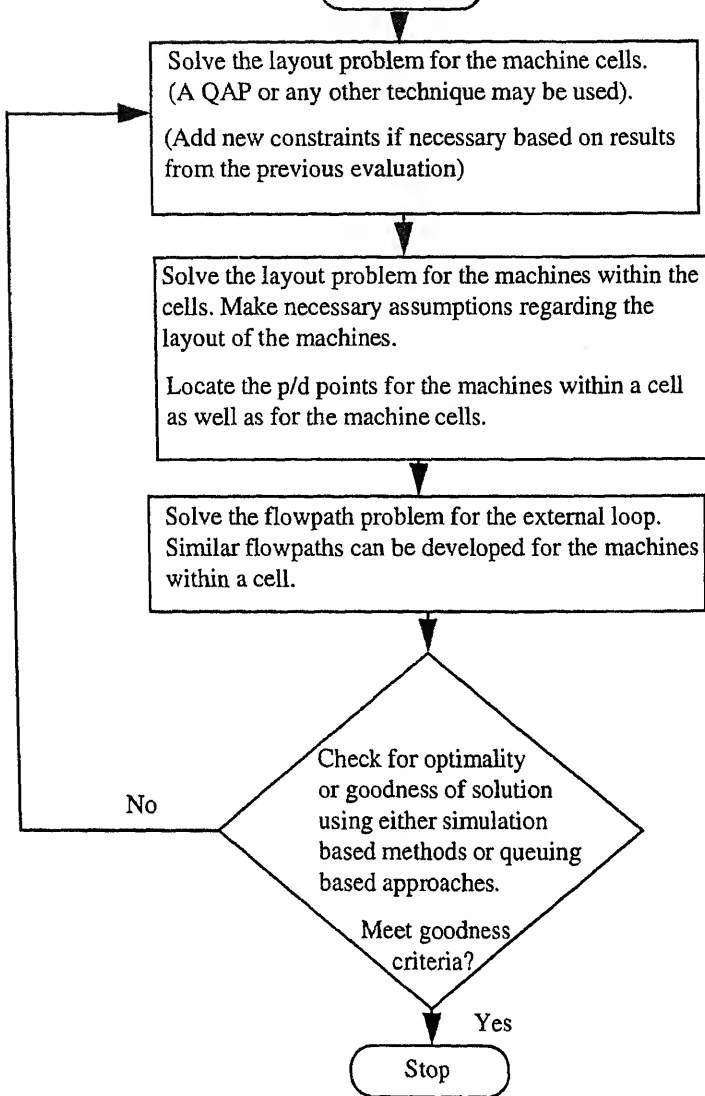
Figure 3. A flowchart describing the methodology used for the integrated design process.

the principles of cellular manufacturing system. The machines are grouped into cells and both the machine layout within a cell and the layout of the cells are decided. The material-handling system for intra-cell material flow is decided based on the machines inside the cell and the operations being performed (figure 4). The P/D points for the cells are also determined.

The layout and the P/D point location can be combined if necessary by using models similar to those developed by Chajjed *et al* (1992). The second stage involves solving the flowpath problem for the cells based on the layout and the P/D points developed. The next stage which is the most important stage in this process involves evaluating the performance
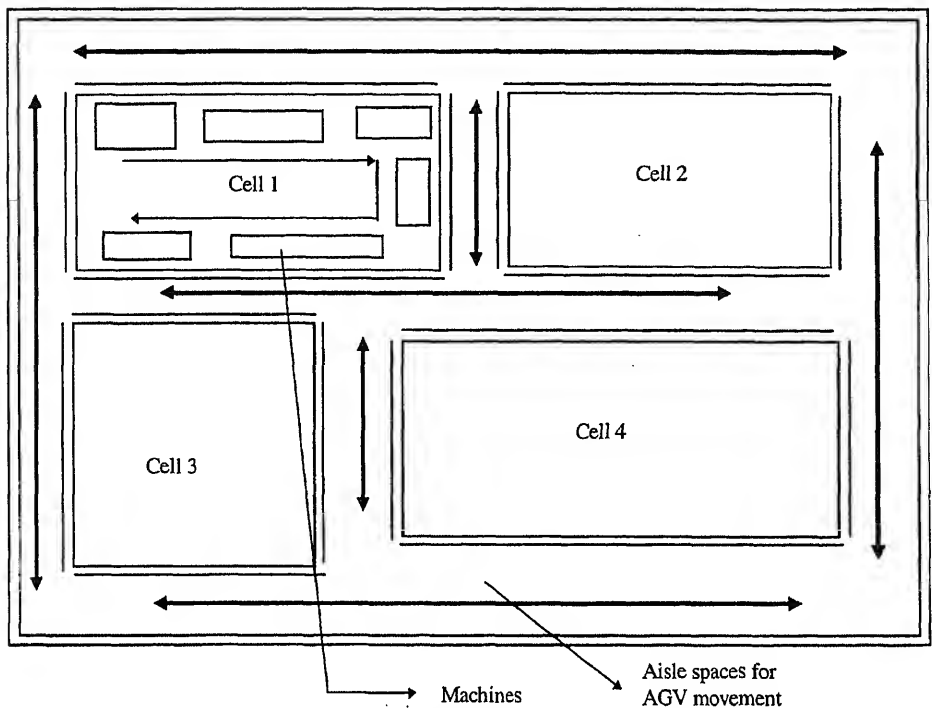
Figure 4. Schematic to show inter cell and intra cell material handling movement.

of the manufacturing system developed based on a predefined set of criteria. A feedback loop is required to ensure that the results obtained from the third stage are communicated to the first stage. The layout is then suitably altered and the process continues until all the criteria are met. The third stage which calculates the "goodness" of a solution consists of either a queuing-based model or a simulation model.

6. Problem formulation

In the case of bi-directional AGVs (where the AGVs can travel in both directions), the flowpath design problem can be formulated as an undirected graph network problem. The P/D points are treated as nodes and the possible flowpaths connecting two P/D points are modelled as arcs. The problem of developing a flowpath is now equivalent to finding the shortest path between any two nodes. The assumptions made in this analysis are:

(1) aisle spaces are known and AGV guides are laid along the aisles;

(2) material transfer rate (or the from – to matrix) is known;

(3) number of AGVs required is already known.

The objective function of the problem can be changed to minimizing the total guidewire length that is to be laid and the problem can be solved as a minimum spanning tree problem. However, this approach assumes that the P/D points are already known. In order to solve the P/D point location and the flowpath problem concurrently, one must assume that the

P/D points are not known. This results in a situation where travel distance between any two nodes is stochastic and the resulting problem is one of stochastic single period network flow.

Every single point along the edge of a cell is a feasible P/D point and therefore in order to determine the optimal or the best P/D points all the points would have to be investigated. This results in a problem with infinite potential points making the problem intractable. For the purpose of this analysis, we will assume that both the pick-up and drop-off points for a given cell are at the same location. This assumption can be relaxed if necessary without increasing the complexity of the problem.

The large continuous problem can be reduced by considering discrete points that represent all the good solution points that exist. In selecting these scenarios, we must also ensure that we do not overlook a "good" or optimal P/D point. A few sample points are selected for each cell such that the maximum deviation from an optimal solution is reduced. The points selected for any cell are the extreme points, points of intersection with other cells, the centre of each side. Although these points may not be representative of the optimal solution to the problem, they will ensure that a good solution is obtained quickly. Hence, for any cell we only consider these 3 sets of points along each edge. In practice, a P/D point is never located at an edge and therefore the P/D points are assumed to be located a small distance 'α' away from the edge, where $\alpha = ((\text{length of P/D station})/2) + \epsilon$ where ϵ is a small value. This limits the number of scenarios we need to consider to a maximum of 17 for each cell. This can be further reduced by using dominance rules.

The algorithm developed by Chajjed *et al* (1992) also helps in selecting P/D points and a flowpath for bi-directional AGVs. However, they assume a rectilinear distance matrix which is valid only in case all the P/D points are situated at extreme points and are of equal sizes. In case of a layout with unequal machines and where the P/D points are not situated at the extreme points of an edge, then their model can underestimate the actual travel distance by as much as the length of the largest side of a cell. For the case of large cells and a large number of trips, this underestimation adds up and results in a significantly lower travel distance than the actual distance. Kim & Klein (1996) also developed an optimal algorithm to develop the P/D points for an AGV system. They prove that the optimal P/D point will always lie at the extreme end of a cell. However, their optimal algorithm and proof is valid only for the case where all the cells or machines are of equal size.

The problem of concurrently determining the P/D points and flowpath in case of bi-directional AGVs can be formulated as a k shortest path network flow problem. In case of small or medium-sized problems, any existing network flow algorithm such as Floyd's algorithm can be used for finding all shortest pairs (Ahuja *et al* 1993). In this case it is assumed that there exist no arcs between the nodes in the same cell to ensure that no more than one node is selected from each cell.

7.   Conclusions

This paper presents an outline of the trends seen in layout and material-handling flowpath

towards more practical ways of tackling these problems and solving them quickly. The research focus is becoming more application based and several problems faced in the implementation stage are considered in the design stage itself. The manufacturing system design is now being considered as a whole instead of splitting it into various sub-problems. Although the entire problem being solved at once may not be feasible, a greater effort is being made to model the entire system or a set of subsystems. This paper also presents an idea for solving the P/D point location and material-handling flowpath problem efficiently and quickly.

References

Ahuja R K, Magnanti T L, Orlin J B 1993 Network flows – theory, algorithms and applications (Englewood Cliffs, NJ: Prentice Hall)

Banerjee P, Zhou Y 1995 Facilities layout design optimization with single loop material flow path configuration. *Int. J. Product. Res.* 33: 183–203

Blazewicz J, Burkard R E, Finke G, Woeginger G J 1994 Vehicle scheduling in two-cycle flexible manufacturing systems. *Math. Comput. Model.* 20(2): 19–31

Bozer Y A, Srinivasan M M 1991 Tandem configurations for automated guided vehicle systems and the analysis of single-vehicle loops. *IIE Trans.* 23: 72–82

Chajjed D, Montreuil B, Lowe T J 1992 Flow network design for manufacturing systems layout. *Eur. J. Oper. Res.* 57: 145–161

Chen J 1995 Cellular manufacturing system design: optimal solution of large scale real world problems. Doctoral dissertation (unpublished), Rensselaer Polytechnic Institute, Troy NY

Dhouib K, Daoud A K 1994 Expert system for AGV managing in bidirectional networks: KADS methodology based approach. *Int. J. Product. Econ.* 33: 31–43

Dowsland K A, Greaves A M 1994 Collision avoidance in bidirectional AGV systems. *J. Oper. Res. Soc.* 45: 817–826

Egbelu P J 1990 Machining and material flow system design for minimum cost production. *Int. J. Product. Res.* 28: 353–368

Egbelu P J, Tanchoco J M A 1986 Potentials for bi-directional guide path for automated guided vehicle based systems. *Int. J. Product. Res.* 24: 1075–1097

Faraji M, Batta R 1994 Forming cells to eliminate vehicle interference and system locking in an AGVS. *Int. J. Product. Res.* 32

Hassan M M D 1994 Machine layout problem in modern manufacturing facilities. *Int. J. Product. Res.* 32: 2559–2584

Heragu S S 1990 Modelling the machine layout problem. *Comput. Ind. Engg.* 19: 294–298

Heragu S S 1992 Recent models and techniques for solving the layout problem. *Eur. J. Oper. Res.* 57: 136–144

Heragu S S 1994 Group technology and cellular manufacturing. *IEEE Trans. Syst., Man Cybern.* 24: 203–215

Heragu S S 1997 Facilities design (Boston, MA: PWS Publishing Company)

Heragu S S, Kusiak A 1991 Efficient models for the facility layout problem. *Eur. J. Oper. Res.* 53: 1–13

Heragu S S, Rajagopalan S 1996 A literature survey of the AGV flowpath design problem. DSES Technical Report # 37-96-446, Rensselaer Polytechnic Institute, Troy, NY

Johnson M E, Brandeau M L 1995 Designing multiple-load automated guided vehicle systems for delivering material from a central depot. *J. Engg. Ind., Trans. ASME* 117: 33–41

Kim J, Klein C M 1996 Location of departmental pickup and delivery points for an AGV system. *Int. J. Product. Res.* 34: 407–420

Kouvelis P, Guitierrez G J, Chiang W C 1992 Heuristic unidirectional flowpath design approaches for automated guided vehicle systems. *Int. J. Product. Res.* 30: 1327–1351

Krishnamurthy N N, Batta R, Karwan M H 1993 Developing conflict free routes for automated guided vehicles. *Oper. Res.* 41: 1077–1090

Kusiak A, Heragu S S 1987 The facility layout problem. *Eur. J. Oper. Res.* 29: 229–251

Maxwell W L, Muckstadt J A 1982 Design of automated guided vehicle systems. *IIE Trans.* 14: 114–124

Noble J S, Klein C M, Midha A 1994 Integrated analysis of material flow systems. Proceedings of the 1994 International Mechanical Engineering Congress and Exposition, MHD 2, pp 33–38

Palliyil G, Goetschalckx M 1996 A comprehensive model for the concurrent determination of aisles and load stations for aisle based material handling systems. Working paper, School of Industrial and Systems Engineering, Georgia Institute of Technology

Peters B A, Smith J S, Venkatesh S 1995 A control classification of automated guided vehicle systems. *Int. J. Ind. Engg.* (accepted)

Sahni S, Gonzalez T 1976 P-complete approximation problem. *J. Assoc. Comput. Mach.* 23: 555–565

Sinreich D 1996 Network design models for discrete material flow systems: A literature review. *Int. J. Adv. Manuf. Technol.* 10: 277–291

Sinreich D, Tanchoco J M A 1991 The centroid projection method for locating pick-up and delivery stations in single-loop AGV systems. *J. Manuf. Syst.* 11: 297–307

Sinreich D, Tanchoco J M A 1993 Solution methods for the mathematical models of single loop AGV systems. *Int. J. Product. Res.* 31: 705–725

Tanchoco J M A, Sinreich D 1992 OSL – optimal single loop guide paths for AGVS. *Int. J. Product. Res.* 30: 665–681

Ulsoy G, Bilge U 1993 Simultaneous scheduling of machines and automated guided vehicles. *Int. J. Product. Res.* 31: 2857–2873

Usher J S, Evans G W, Wilhelm M R 1988 AGV flow path design and load transfer point location. IIE International Industrial Engineering Conference Proceedings, pp 174–179

Wang H P B, Hafeez S A 1994 Performance evaluation of tandem and conventional AGV systems using generalized stochastic petri nets. *Int. J. Product. Res.* 32: 917–932

Wu C T, Egbelu P J 1994 Concurrent design of shop layout and material handling. Working paper, Department of Industrial and Management Systems Engineering, Pennsylvania State University

Recent Advances in Mechanical Engineering

Foreword

This special issue of *Sādhanā* is a collection of nine invited lectures presented at the International Conference on Advances in Mechanical Engineering (ICAME), held at the Indian Institute of Science, Bangalore in December 1995, to commemorate the Golden Jubilee of the setting up of the Department of Mechanical Engineering. In these fifty years the department has played a significant role in the building up of a sound technological base in the country through the active interaction of its faculty and alumni with major industries, national laboratories and other research and development institutions in India and abroad, and has stimulated a continuous quest for challenging new directions in teaching, research and development activities. The department appropriately celebrated its golden jubilee by organising the ICAME to take stock of recent developments and envision future trends in mechanical engineering. When a proposal was made to disseminate the invited ICAME talks to a wider audience of scientists and engineers through *Sādhanā*, it met with ready encouragement from the Indian Academy of Sciences, and in particular from Professors N Viswanadham and J Srinivasan, for which I thank them.

The nine papers in the present issue represent a broad spectrum of current research in mechanical engineering. The first paper by Abhijit Guha considers two-phase flow of vapour–liquid mixtures with large numbers of minute droplets. He analyses the formation of and the impact of these droplets on the dynamics of this flow. Two-phase flows are present in a large number of industrial and natural situations, such as dusty gases, cavitation, steam turbines etc. Guha proposes a new theory to explain the nucleation process in a steam turbine.

The second paper by Fazle Hussain and others describes vortex mechanics based on the unique characteristics of confined swirling incompressible flows. The development of the vortex engine is described in detail. Experiments and analyses performed on a laboratory "cold" model of the vortex engine are presented.

Noncircular exits for gas-jets have been explored as a passive and inexpensive method to control the characteristics of fuel gas-jets, thereby controlling the combustion and pollutant characteristics. S R Gollahalli's paper describes experimental studies on noncircular jet flames including flow visualisation of the structure of the jets.

In the field of power transmission, compactness of the transmission element is becoming increasingly important. While stress and strength considerations have played a dominant role in design, controlling the load factor through load distribution/equalisation can be a powerful tool to realise compact transmissions. K Lakshminarayana's paper presents the issues important in design for effective load distribution in transmission elements, in

particular, spur and helical gear systems. Application of parallel power paths to achieve better load distribution is explained.

The fifth paper in the issue by K N Gupta discusses the analysis of vibration for condition monitoring and diagnosis, including basic principles, available instrumentation and various vibration monitoring techniques. The paper also presents interesting case studies that use vibration signatures for diagnosing failures.

The paper by Subhash Sinha presents a general technique for the analysis of time-periodic nonlinear dynamical systems. In his approach, Sinha uses the Liapunov–Floquet transformation to convert quasi-linear periodic systems to linear systems that are time-invariant. The paper illustrates stability and bifurcation analyses of the transformed systems with two examples.

The seventh paper by Amitabha Ghosh describes an experimental investigation on the non-traditional manufacturing process of electrochemical discharge machining. The author studies the process of spark generation and identifies inductance in the circuit as the significant process parameter. Two innovative applications of the process for micro-welding and rapid-prototyping are also proposed.

The eighth paper by M Ramulu investigates machining mechanisms and machining-induced effects on the structural integrity of FRP materials. The paper presents results of experimental studies of orthogonal cutting of both unidirectional and multidirectional graphite/epoxy composites.

The process of removal of material in the machining and surface finishing of brittle solids, such as ceramics, is studied by S Chandrasekar and T N Farris in the last paper of the issue. They present a model of the material removal process and identify the mechanisms responsible for material removal. The effect of these mechanisms on the machined surface is also described.

Though the papers in the issue have not covered all areas of the vast field of mechanical engineering, they present a sample of the work going on around the world. I hope the issue will be of interest to researchers and practising engineers.

I would like to end this foreword with an acknowledgement of the untiring efforts of Ms Shashikala and other editorial staff of the Academy in bringing out this special issue, an expression of sincere thanks to the contributors for their ready cooperation and a personal remembrance of one of the contributors, Prof K Lakshminarayana, who is sadly not in our midst today to see his lecture in print in *Sādhanā*.

June 1997   T S MRUTHYUNJAYA
                                                                        Guest Editor

Analysis and computation of non-equilibrium two-phase flows

ABHIJIT GUHA

Department of Aerospace Engineering, University of Bristol, Queen's Building, University Walk, Bristol BS8 1TR, UK
e-mail: A.Guha@Bristol.ac.uk

Abstract. Non-equilibrium fluid mechanics and thermodynamics of two-phase vapour-droplet and gas-particle flow are considered. Formation of the droplets as well as their subsequent interaction with the vapour are discussed. A new theory of nucleation in steam turbines is developed that reproduces many aspects of measured droplet size spectra which cannot be explained by any available steady-flow theories. (Steam turbines are responsible for 80% of global electricity production and the presence of moisture significantly reduces turbine efficiency costing 50 million pounds per annum in UK alone.) Fluid dynamic interactions discussed include flow instabilities induced by condensation, condensation wave theory, relaxation gas dynamics for vapour-droplet flow, thermal choking due to non-equilibrium condensation, the structure of shock waves and their development through unsteady processes, and jump conditions and the interpretation of total pressure in two-phase flows.

Keywords. Non-equilibrium fluid mechanics; thermodynamics of two-phase flows; nucleation in steam turbines.

1. Introduction

Two-phase flow of a vapour–liquid mixture consisting of a large number of minute liquid droplets uniformly dispersed throughout a background vapour phase continuum is both scientifically interesting and of considerable engineering importance (in a variety of areas in mechanical, chemical and aerospace engineering, and meteorology). In the present paper, we discuss the formation of the droplets as well as their impact on the subsequent thermo-fluid dynamics of the flow. We restrict ourselves mainly to the description of *pure* substances, i.e., when both phases are of the same chemically pure species. Numerical illustrations are given only for steam–water mixtures, but the general principles are applicable to other substances as well.

What follows is a brief description, with no equations, of some of the work with which the present author has been involved over the past few years. The references cited (Guha

& Young 1989, 1991, 1994; Young & Guha 1991; Guha 1992, 1994, 1997) give a fuller treatment of these topics. The VKI lecture series (Guha 1995) treats most of these topics at much greater depth. It also contains a good repertoire of references and works of many researchers in this field which we do not reproduce here owing to space constraints.

2. Thermo-fluid dynamics of condensation

2.1 *Physical description of homogeneous condensation*

All condensing (or evaporating) flows are non-equilibrium to a greater or less extent. Departures from equilibrium are measured by the subcooling ΔT which is the difference between the saturation temperature at local pressure and the actual vapour temperature ($\Delta T = T_s - T_g$). ΔT governs the rate at which nuclei are formed as well as the rate at which established droplets grow (or evaporate).

As pure, clean steam expands through a nozzle or a turbine blade passage, droplets do not appear as soon as the condition line crosses the saturation line. This is due to the existence of a free-energy barrier involved in creating new surface area. For some considerable time during expansion the steam remains dry in a metastable equilibrium until the subcooling becomes high enough to trigger an appreciable nucleation rate. Depending on the rate of expansion and the pressure, steam may become subcooled by 30–40°C while still remaining dry. The nucleation process leads to the formation of very large number (10^{14}–10^{17} nuclei per kg of steam) of tiny droplets (diameter $<$ 1 nm), called the primary fog, more or less uniformly distributed in the continuous vapour phase. Nucleation is practically terminated at the point of maximum subcooling called the Wilson Point. For pure steam, if the Wilson points for tests with varying nozzle inlet conditions are plotted on the equilibrium Mollier diagram, they are contained within a narrow zone around a line called the Wilson line (which corresponds to approximately 3–4% equilibrium wetness line).

The droplets thus formed then *rapidly* grow in size by exchanging heat and mass with the surrounding, subcooled vapour (the final droplet radii, r, in laboratory nozzles lie in the range 0.02–0.2 μm). The high rate of heat release as a result of rapid condensation, causes a sharp increase in vapour temperature and consequently an exponential decay of the subcooling. Depending on the values of the flow parameters, the initial growth phase of the droplets may give rise to a gradual increase in pressure known as "condensation shock". The term "shock", however, is a misnomer. Although pressure rises as a result of heat addition to supersonic flow, the Mach number downstream of the condensation zone usually remains above unity and, more importantly, the rise in pressure is gradual.

In conventional laboratory nozzle experiments, where (dry saturated or superheated) steam is produced in a boiler, the flow must expand to supersonic velocities for significant subcooling to develop. However, if subcooled steam could be supplied at the nozzle inlet, homogeneous condensation could occur in the subsonic part of the flow. This is possible, for example, in a multistage turbine where steam could become subcooled at the inlet of

densation zone ("supercritical condensation shock"). Under certain conditions this shock wave may become unstable and propagate towards the nozzle throat. The compressive wave ultimately interferes with the nucleation zone causing a reduction in nucleation rate and hence heat release rate. With the cause of its inception removed, the strength of the wave decreases and the flow again expands through the throat in a shock-free manner thus allowing the whole process to repeat itself. Such unsteady flow is normally observed in pure steam when the inlet stagnation temperature T_0 is close to the saturation temperature at the inlet stagnation pressure p_0. Homogeneous condensation then occurs in the transonic region close to the throat causing flow instability (the flow domain and boundary conditions remaining fixed).

Keeping p_0 fixed, if T_0 is progressively reduced from superheated to subcooled levels, one encounters different regimes of homogeneous condensation in the order: Subcritical condensation (the usual pressure humps characteristic of many condensation experiments), supercritical condensation (with inbuilt shock wave), oscillatory condensation and subsonic condensation (Guha 1994a).

After the "condensation shock", the steam generally reverts to near thermodynamic equilibrium at which the temperature of the vapour as well as of the droplets is close to the saturation level. Since the growth of liquid phase takes place by heat transfer through a finite temperature difference between the phases, the process is essentially irreversible and has associated with it a net rise in entropy. In turbines this appears as a reduction in the potential for performing work and is referred to as the *thermodynamic wetness loss*. This is a major component of the overall wetness loss. A simplistic version of an empirical rule, formulated by Baumann in 1921, states that the efficiency of a steam turbine decreases by 1% for every 1% increase in mean wetness fraction. A typical value of the wetness fraction at the exit of a steam turbine in an electricity-generating power-plant is 10–12%. Thus, in the last stages, the wetness loss is comparable to the combined effects of the profile, secondary and tip leakage losses. A 1988 estimate by the then Central Electricity Generating Board of UK showed that the adverse wet steam effects cost them 50 million pounds per year. The global implication is thus quite serious, since steam turbines are responsible for about 80% of the world-wide generation of electricity and there is a considerable economic incentive for further research.

2.2 *Numerical solutions for different regimes of condensation*

The numerical scheme for the calculation of steady as well as unsteady non-equilibrium wet steam flow has been detailed in Guha & Young (1991) and Guha (1994a). Here, we describe only the outline and highlight some important aspects.

One of the most effective methods of calculation is to write a computational "black-box" which contains the nucleation and droplet growth equations, and the energy equation in its thermodynamic form. Together they furnish the full set of equations that describe completely the formation and growth of liquid droplets in a fluid particle (from a Lagrangian viewpoint), if the pressure–time variation is specified. The pressure–time variation is obtained by time-marching solutions of the conservation equations such as Denton's method,

extensively used for single-phase calculations in turbomachinery blade rows. In this respect, the thermodynamic aspects of phase-change can be completely divorced from fluid dynamical considerations so that the use of the "black-box" is effectively independent of any particular computational fluid dynamic application. Thus established single-phase CFD codes can, rather easily, be modified to deal with non-equilibrium two-phase flow with the above-mentioned *modular approach*. (The flexibility of this scheme may be appreciated from Guha & Young (1994) where the same "black-box" has been grafted into a streamline curvature calculation procedure.)

The development of the computational routines within the "black-box" represents a comparatively major undertaking and has been fully described by Guha & Young (1991) and Guha (1994a). The routines are sufficiently general and robust to deal with any type of nucleating or wet steam flow and (in contrast to many procedures reported in the literature) *full details of the polydispersed droplet size spectrum following nucleation are retained in the calculations. The last aspect is essential for accurate modelling of the nucleation zone.* This has been possible, without consuming excessive CPU time, by developing a novel averaging procedure that constantly redefines the average size and droplet number in each droplet group. In this way, the number of droplet groups required is restricted to an affordable optimum, while always retaining the correct shape of the droplet size spectrum.

A mixed Eulerian–Lagrangian technique is used. The continuity and momentum equations are solved by Denton's time-marching method. (The wet-steam "black-box" being flexible and modular, any other Eulerian time-marching method, e.g. Jameson's scheme, can be used.) The "black-box" performs the integration of the droplet growth equations along the fluid path lines rather than the usual, quasi-unsteady, method in which the pressure field remains frozen at a given instant of time while the growth of the liquid phase is calculated. *The present scheme allows simultaneous solution of all the relevant equations and enforces the correct coupling between the vapour-phase gasdynamics and the relaxation effects due to the presence of the liquid phase.*

For a proper comparison between experiments and theory, variation in pressure as well as droplet size must be considered. Many references compare the variation in pressure only. Such comparison is an inadequate test for nucleation and droplet growth theories. Almost any nucleation theory can be 'tuned' to reproduce the measured pressure distribution. A crucial test is to find out whether the same 'tuning' can predict the correct droplet size as well. Experience with calculations for wet steam points out categorically that, in general, predicting a satisfactory pressure distribution does not automatically ensure a good prediction of droplet size. The present computational scheme has been validated against measurements of steady (both sub- and supercritical) and unsteady condensation shock waves (Guha & Young 1991). Various regimes of condensation have been computed by Guha (1994a), which shows a novel example of subsonic condensation where the nozzle is unchoked at the geometric throat. (In most reported studies on condensation shocks in nozzles, condensation takes place in the supersonic divergent part.)

Figure 1 presents one example of unsteady calculation. The prediction compares well with measurement. The pressure profiles at different instants during a complete cycle reveal exactly the same sequence of the formation and movement of the shock wave as explained earlier. As the aerodynamic shock wave moves upstream towards the throat and interacts with the nucleation zone, progressively fewer droplets are nucleated, thus resulting in a
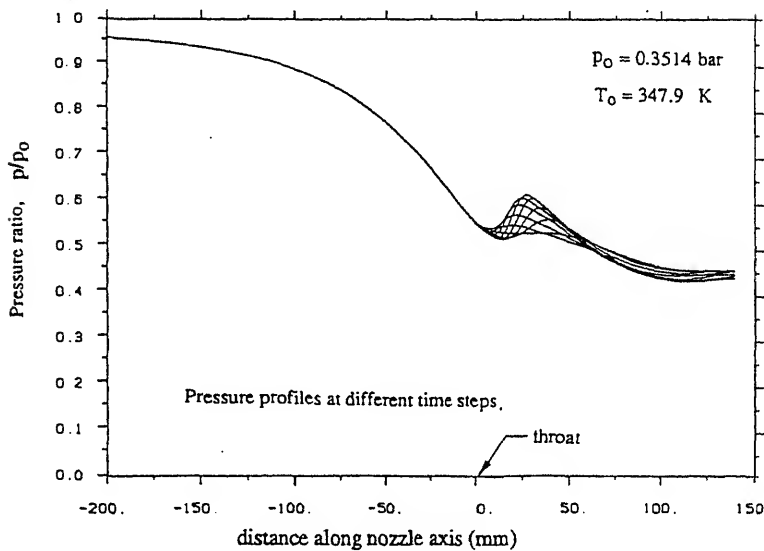
Figure 1. Unsteady condensation: Evolution of pressure for one cycle.

larger final mean radius. This causes a large variation in the droplet sizes during each cycle. An interesting implication of the unsteady nucleation process is that it may be a contributing factor to the formation of the highly-skewed polydispersed droplet spectrum measured in a real steam turbine: a polydispersity which cannot be predicted with steady flow calculation methods (see § 4). Details of this and other calculations are in Guha & Young (1991) and Guha (1994a).

2.3 *Integral analysis: Condensation wave theory*

A great deal of physics may be learnt from an integral analysis, which is a study of the jump conditions relating the end states of the condensation zone, without considering the detailed flow structure within it. Figure 2 shows an example calculation of the condensation wave theory for air with 1% moisture content. Guha (1994a) presents condensation wave
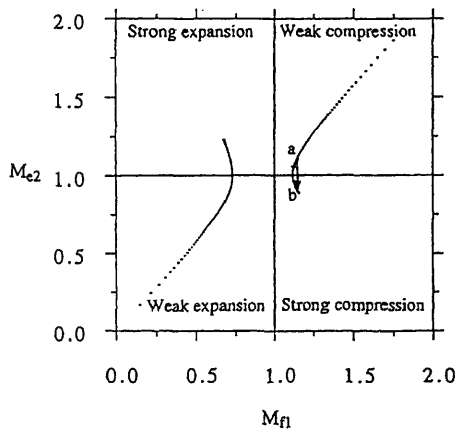
Figure 2. Condensation wave theory.

theory in great detail and discusses similarities and differences with the more familiar combustion wave theory. (The reference considers control volume analysis of two types of near-discontinuities, condensation shock as well as aerodynamic shock waves in vapour-droplet flow, from a common standpoint.) In figure 2, M_{f1} represents the upstream Mach number based on the frozen speed of sound a_f and M_{e2} represents the downstream Mach number based on the equilibrium speed of sound a_e (§ 3.2). The effect of heat release is very pronounced when the Mach number lies approximately between 0.8 and 1.1, and a small amount of condensation may alter the flow velocity etc., quite dramatically. Details may be found in Guha (1994a).

2.4 *Thermal choking due to non-equilibrium condensation*

It is well known that heat addition causes a reduction in Mach number in supersonic flow and an increase in Mach number in subsonic flow. In other words, heat addition to a flowing fluid drives the Mach number towards unity. Therefore, at a particular flow Mach number, the fluid can absorb a maximum quantity of heat before the local Mach number equals unity and the flow becomes thermally choked. Any elementary textbook on classical gas dynamics gives the expression for critical quantity of heat, $q_{classical}^*$, for simple heat addition to an ideal gas (external heat addition without any change in flow cross-sectional area).

Similar to the case of external heat addition, the Mach number decreases in the condensation zone (the flow being supersonic). Therefore, for particular combinations of nozzle geometry, supply conditions and the working fluid, the liberation of latent heat could be such that the minimum Mach number becomes unity and the flow is thermally choked. (A numerical computation of this limiting case of thermal choking due to non-equilibrium condensation is later shown in figure 4.) If the inlet total temperature, T_{o1}, is reduced any further, keeping the inlet total pressure, p_{o1}, fixed, continuous variation of the flow variables is no longer possible and an aerodynamic shock wave appears inside the condensation zone.

Although widely referred, it has been argued in Guha (1994b) that the expression for $q_{classical}^*$ is not appropriate for a condensing flow primarily for two reasons: (i) In case of a condensation shock, the heat is added as a result of condensation of a part of the fluid itself. Therefore, the mass flow rate of the condensable vapour changes as the vapour is continually transformed into the liquid phase. The expression for $q_{classical}^*$, which is derived for external heat addition to an ideal gas, does not take into account this mass depletion. (ii) The droplets formed through homogeneous nucleation grow at a *finite rate* by exchanging mass and energy with the surrounding vapour. Therefore, the energy addition due to condensation is not instantaneous and takes place over a short but finite zone. Since condensation shock normally occurs in the diverging section (with dry vapour at inlet), this means the flow area increases between the upstream (A_1) and downstream (A_2) of the condensation zone. The expression for $q_{classical}^*$, on the other hand, is derived by assuming heat addition in a constant area duct.

Guha (1994a) gives details about the relative importance of the above two effects. Two new expressions for critical quantity of heat have been derived: q_A^* which deals with external heat addition with area variation, and $q_{integral}^*$ which takes into account the area variation as well as the depletion in the mass of vapour due to condensation. The dotted
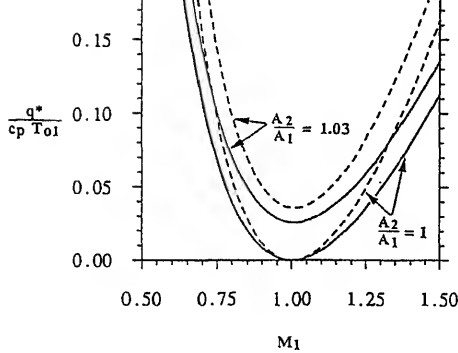
Figure 3. Effects of area ratio across the condensation zone and the depletion in vapour mass flow rate on the critical amount of condensation ($— q_A^*$, $- - - q_{integral}^*$).

lines in figure 3 represent $q_{integral}^*$ while the solid lines are the plots of q_A^*. Note that the solid line corresponding to unity area ratio ($A_2/A_1 = 1$) is nothing but the classical solution, $q_{classical}^*$. Figure 3 shows clearly that the effects of mass depletion and (even a small) area variation are quite dramatic, especially when the Mach number is close to unity.

Figure 4 presents a time-marching solution of a limiting condensation shock, giving the variation in Mach number (M), non-dimensional pressure (p/p_{o1}) and non-dimensional wetness fraction (y/y_{eq}) along the nozzle axis. There are two sonic points in the flow field. (Achieving exactly a Mach number of unity at the end of the condensation zone (point 2) required numerical experiments of varying T_{o1} keeping p_{o1} fixed.) At the conditions shown in figure 4 $A_2/A_1 = 1.0284$, $M_1 = 1.182$. The critical amount of heat obtained from this full numerical solution of the nozzle flow is denoted by q_{actual}^*. Table 1 compares the results of various integral analyses.

In conclusion, a theory of thermal choking due to non-equilibrium condensation in a nozzle is presented (Guha 1994b). The theory is based on a simple control volume approach. (A differential theory of thermal choking is discussed in Guha (1994a).) It applies to vapour-droplet flow with or without a carrier gas. The expression for critical heat (or condensation) derived is valid for either supersonic or subsonic flow, and for heat release either in the diverging or in the converging part of a nozzle.
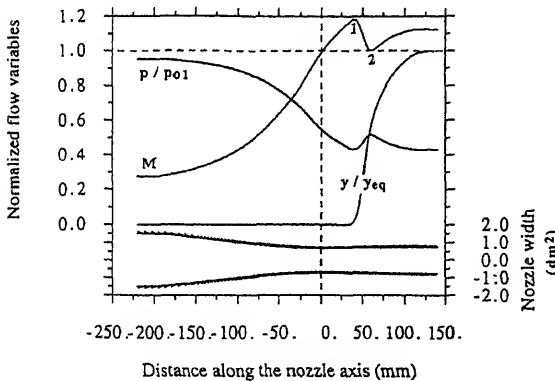
Figure 4. Time-marching solution of a limiting condensation shock in a quasi-one-dimensional convergent-divergent nozzle. (*Continuous* variation of flow variables leading to thermal choking at point 2.)

Table 1. Comparison of integral predictions with exact solution.

External heat addition to ideal gas		Condensational heat release with area variation and mass depletion of vapour $(q^*_{integral}/c_{pg}T_{o1})$	Time-marching solution of non-equilibrium gas dynamic equations $(q^*_{actual}/c_{pg}T_{o1})$
in constant area duct $(q^*_{classical}/c_{pg}T_{o1})$	in duct of varying area $(q^*_A/c_{pg}T_{o1})$		
0.01978	0.0423	0.0574	0.0587

Table 1 shows that the present theory, $q^*_{integral}$, is in very close agreement with the full numerical solution of the differential equations of motion (giving the detailed structure of a condensation shock wave leading to thermal choking). The usually quoted $q^*_{classical}$ underestimates the critical heat by a factor of three in the example calculation presented. The variation of area across the condensation zone (although small) and the depletion in vapour mass as a result of condensation cannot be neglected in determining the critical heat in condensing nozzle flow.

3. Fluid dynamics with interphase transport of mass, momentum and energy in pure vapour-droplet mixtures

3.1 *Relaxation gas dynamics for vapour-droplet mixtures*

A lucid description covering many aspects of relaxation gas dynamics and its applications to vapour-droplet flows (including coupled relaxation processes) may be found in Guha (1995).

If a property of a medium is perturbed from its equilibrium state and the restoration to equilibrium occurs at a *finite rate*, the medium is called a relaxing medium and the process of restoration is termed relaxation. Simple relaxing media follow the archetypal rule of the restoration process following a perturbation: $d(\Delta q)/dt = -\Delta q/\tau$, where, q is an internal state variable, Δq is the departure from equilibrium and τ is the relaxation time.

A vapour-droplet medium is assumed to be a homogeneous two-phase mixture of a large number of fine, spherical droplets dispersed in the continuous vapour phase. Although the droplet cloud may exhibit an arbitrary level of polydispersity with a spectrum of different sizes of droplets, in the present analysis we will restrict ourselves to a mono-dispersed droplet population for simplicity of description. We also consider pure substances only. The droplets are large so that the capillary subcooling is negligible.

A vapour-droplet medium may go out of equilibrium in three different ways as below.

(i) The droplet temperature is not equal to the saturation temperature ($T_l \neq T_s$). $\Delta T_l = T_s - T_l$ is the relevant non-equilibrium variable and τ_D is the corresponding droplet temperature relaxation time.

(ii) The two phases have unequal velocities, i.e., there may be a slip between the phases ($V_g \neq V_l$). $\Delta V = V_g - V_l$ is the non-equilibrium variable and τ_I is the corresponding inertial relaxation time.
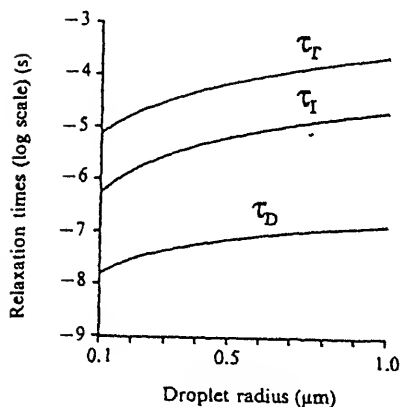
Figure 5. Relaxation times for water droplets in pure steam.

(iii) The temperature of the vapour phase may be different from the saturation temperature ($T_g \neq T_s$). $\Delta T = T_s - T_g$ is the relevant non-equilibrium variable and τ_T is the corresponding vapour thermal relaxation time. [Note that negative ΔT means that the vapour is superheated.]

It is clear that under non-equilibrium conditions mass, momentum and energy transfers would take place between the two phases. It is through these interphase transfer processes that the system returns to equilibrium following a perturbation. The equations representing the different interphase transfer processes and the expressions for τ_D, τ_I and τ_T have been derived in Young & Guha (1991) and Guha (1995). Figure 5 shows the variation of the three relaxation times in pure steam ($p = 0.5$ bar, $y = 0.1$) as a function of droplet radius.

It can be shown (Guha 1992c) that, quite generally, the relation $\tau_D \ll \tau_I \ll \tau_T$ holds in pure vapour-droplet flow. Therefore, following a disturbance, on a very short time scale the droplet temperature reaches equilibrium, then the velocity slip and finally the vapour temperature relaxes to the equilibrium value. *While considering any particular relaxation process, therefore, relaxation processes with smaller relaxation times may be assumed to have been equilibrated whereas relaxation processes with larger time scales may be assumed to be frozen.* Mathematically, for example, *equilibrium* droplet temperature means that ΔT_l and τ_D simultaneously tend to zero, the ratio $\Delta T_l/\tau_D$ remaining finite such that $\Delta T_l/\tau_D \rightarrow V_l(dT_s/dx)$, and, for instance, *frozen* heat transfer implies $\tau_T \rightarrow \infty$.

3.2 *Speeds of sound in a pure vapour-droplet mixture*

As a result of three relaxation processes in vapour-droplet flow, four different sound speeds (a_f, a_{e1}, a_{e2}, a_e) may be defined subject to different mechanical and thermodynamic constraints. The full frozen speed of sound, a_f, corresponds to the speed of an harmonic acoustic wave of such high frequency that the response of the droplets is negligible (i.e. zero mass, momentum and energy transfer). The full equilibrium speed of sound, a_e, corresponds to the speed of an harmonic acoustic wave of such low frequency that liquid–vapour equilibrium is maintained at all times. The two intermediate speeds correspond
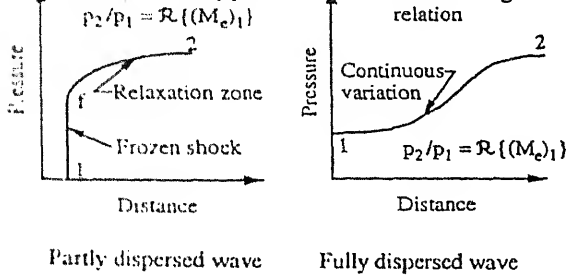
Figure 6. Schematic structure of shock waves. (a) Partly and (b) fully dispersed waves.

heat transfer (a_{e1}), and (b) the case of equilibrium droplet temperature and velocity slip relaxation but frozen heat transfer (a_{e2}).

The derivation and the expressions of the sound speeds may be found in Young & Guha (1991) and Guha (1995). The relationship of the four sound speeds to each other is of great importance. As a typical example, the ratios of different sound speeds for steam at 1 bar and 0.1 wetness fraction are given by, $a_f: a_{e1}: a_{e2}: a_e \equiv 1:0.997:0.945:0.878$.

3.3 Structure of stationary shock waves

One can define at least two Mach numbers corresponding to the limiting speeds of sound: a frozen Mach number $M_f = V/a_f$, and an equilibrium Mach number $M_e = V/a_e$, where V is the flow velocity. In general, $a_e < a_f$ and $M_e > M_f$. As a result of the frequency dispersion, therefore, two distinct types of shock waves might form in a relaxing medium, the far upstream and far downstream conditions being at equilibrium (figure 6): (1) A partly dispersed shock wave, where a discontinuous jump in flow variables (dominated by viscosity and thermal conductivity as in Rankine–Hugoniot shocks in classical gas dynamics) is followed by a continuous relaxation zone in which the fluid returns to equilibrium by performing the relaxation processes. Such a case arises if the upstream velocity is greater than the frozen speed of sound, i.e., if $V_{go} > a_{fo}$ or, $M_{fo} > 1$. In order to denote the upstream condition, we use both subscripts o (so that confusion does not arise with sound speeds a_{e1} and a_{e2}) and 1 (as this is the conventional nomenclature for Rankine–Hugoniot conditions discussed in § 3.5). (2) A fully dispersed shock wave, where a *continuous* change of flow properties takes place from the upstream to the downstream equilibrium state. This corresponds to the case when $a_{fo} \leq V_{go} < a_{eo}$; i.e., when $M_{fo} \leq 1$, but $M_{eo} > 1$.

With four limiting speeds of sound in a vapour-droplet mixture, four types of stationary shock wave structures may arise (Young & Guha 1991; Guha 1992c, 1995):

type I waves	corresponding to	$a_{eo} < V_{go} < a_{e2o}$,
type II waves	corresponding to	$a_{e2o} < V_{go} < a_{elo}$,
type III waves	corresponding to	$a_{elo} < V_{go} < a_{fo}$,
partly dispersed waves	corresponding to	$V_{go} > a_{fo}$.

Type I, II and III waves are sub-categories of the fully dispersed waves, where the steepening effect of the nonlinear terms in the equations of motion is just balanced by the dispersive effect of the relaxation processes. Type I waves are dominated by vapour thermal relaxation, type II waves by both velocity and vapour thermal relaxation and type III waves by all three relaxation processes. It has been shown (Young & Guha 1991; Guha 1995)

that $|\Delta T_l|$ becomes unstable in the interval $a_{e1} < V_g < a_f$, $|\Delta V|$ becomes unstable in the interval $a_{e2} < V_g < a_{e1}$, and, $|\Delta T|$ becomes unstable in the interval $a_e < V_g < a_{e2}$. These instabilities in the non-equilibrium variables are the reasons for the existence of the fully dispersed waves. Figure 7 shows numerical solutions for a typical type II fully dispersed wave in a steady flow of pure wet steam. Numerical solutions of other types of waves may be found in (Guha & Young 1989; Guha 1992b, 1995).

The usual model of a partly dispersed shock wave assumes that the interphase transfer processes are frozen during the passage through the discontinuity and the vapour properties just downstream of the discontinuity can be calculated using a standard Rankine–Hugoniot analysis. The liquid droplets therefore pass through the frozen shock without change in radius, temperature and velocity. The conditions downstream of the discontinuity provide the initial values for integrating the two-phase conservation equations through the relaxation zones. The droplet temperature relaxes very quickly, followed by velocity slip and finally the vapour temperature. The lengths of the relaxation zones are in the approximate ratios $\tau_D : \tau_I : \tau_T$. Details about partly dispersed waves in vapour-droplet flow may be found in (Guha 1992b). Although linearized analyses are often presented for the relaxation zone downstream of the frozen shock in standard treatments on relaxation gas dynamics, they are of limited applicability to vapour-droplet flows (Guha 1992b).

3.4 *Shock waves in unsteady flow*

The physical significance of the various wave profiles discussed above can be appreciated more readily by considering their development under unsteady flow conditions. As a typical example, we now discuss wave generation in one-dimensional flow by an instantaneously accelerated piston in a frictionless pipe initially containing stationary wet steam. The numerical scheme and other details may be found in Guha & Young (1989).

Figure 8 shows the numerical prediction of a wave propagating in wet steam. The figure also includes the flow behaviour in a dry, ideal gas under identical conditions in order to illustrate the special features of a vapour-droplet mixture. The $(t\text{-}x)$ diagram was constructed from the results of the unsteady time-marching calculation. At the instant of initiation, all the interphase transfer processes are frozen and the shock velocity corresponds to the propagation velocity into a single-phase vapour at the same temperature. Behind the shock, the mixture relaxes to equilibrium along the particle pathlines. The droplet temperature relaxes first on the very short timescale τ_D and is followed by the velocity slip and vapour temperature on timescales τ_I and τ_T respectively. Changes along the particle paths are propagated upstream and downstream along the left and right running Mach lines (based on the frozen speed of sound). The right running Mach lines overtake the shock wave, weakening it and causing it to slow down. The shock path therefore curves in the $(t\text{-}x)$ diagram until it reaches a constant equilibrium speed. When this occurs, the dispersive effects of the relaxation processes are just balanced by the steepening effects of the non-linear terms and the wave structure is identical to that of the stationary waves in steady flow described earlier. Whether the final equilibrium structure is partly or fully dispersed
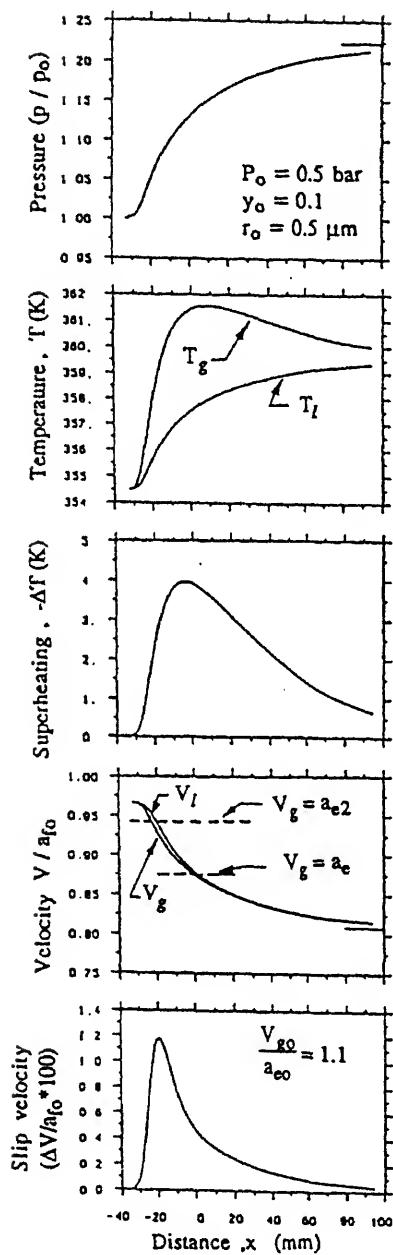
Figure 7. Numerical solution for a Type II wave in wet steam.
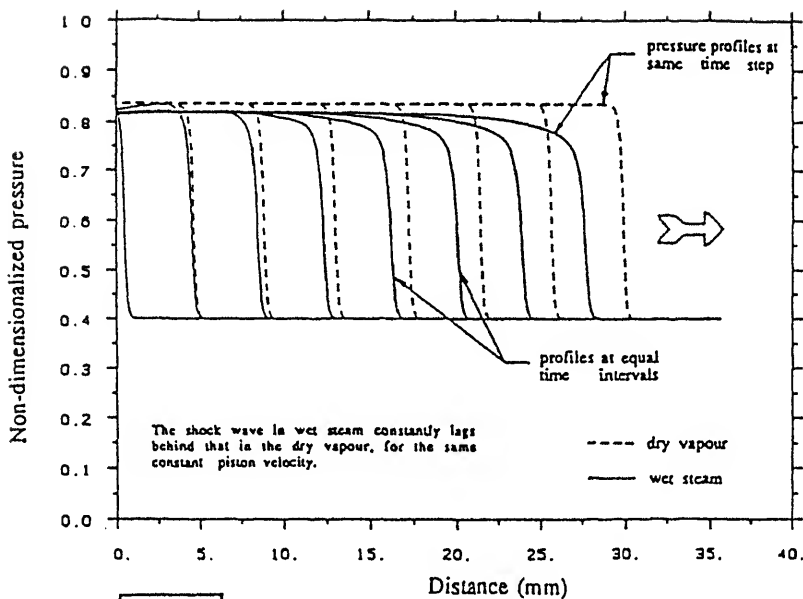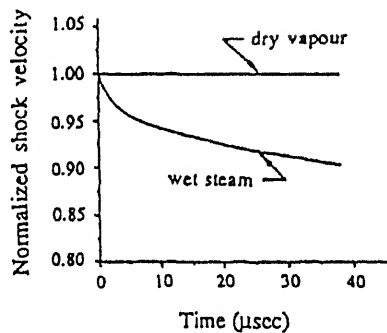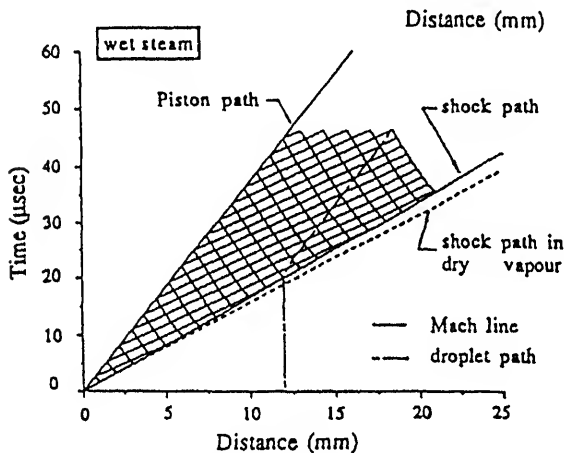
Figure 8. Numerical prediction of shock propagation in a pipe.

superheat vapour temperature, droplet radius and wetness fraction is shown by the curves in figure 9 which are self-explanatory. As with stationary partly dispersed shock waves, the increase in wetness fraction downstream of the frozen shock wave is due to the effects of velocity slip.

3.5 *Integral analysis: Jump conditions*

Detailed study on the jump conditions across shock waves has been made by Guha (1992a, 1994a) and the similarities and differences of condensation discontinuities and aerodynamic shock waves are discussed at length (Guha 1994a). Here, we can only mention the bare minimum.

In a simple relaxing medium, e.g., a solid-particle-laden gas, the form of the jump conditions (R in figure 6) are identical to the Rankine–Hugoniot relations for an ideal gas,
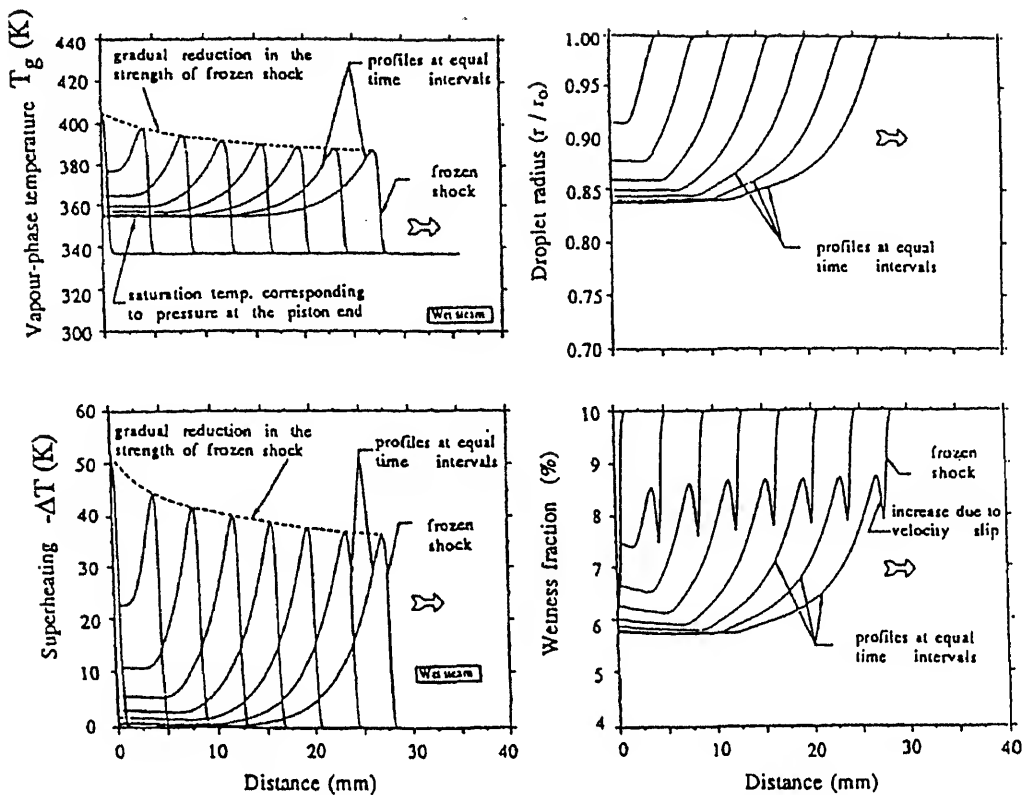
Figure 9. Variation of flow parameters during shock propagation in wet steam.

if one uses the equilibrium Mach number, M_e, and the equilibrium isentropic index, γ_e. (Expressions for γ_e in vapour-droplet mixtures with or without a carrier gas are given in Guha 1992a, 1994a) However, it has been shown by Guha (1992a) that although jump conditions of the same form as Rankine–Hugoniot relations can be formulated for vapour-droplet mixtures, they are *approximate* and hold only *conditionally*.

A solid-particle-laden gas can be treated as a modified gas. Equations governing vapour-droplet flow are much more complex. An additional complexity (and *novelty*) occurs due to interphase mass transfer (the liquid phase *evaporates* inside a dispersed wave). If the strength of the wave is sufficient, complete evaporation may result, thereby rendering a two-phase medium before a shock wave into a single-phase one after it!

The integral analysis (Guha 1992a) reveals that, depending on the upstream wetness fraction and the pressure ratio across the wave, four types of shock structures may result in vapour-droplet flow. They are: (I) equilibrium fully dispersed, (II) equilibrium partly dispersed, (III) fully dispersed with complete evaporation, (IV) partly dispersed with complete evaporation. Figure 10 shows the boundaries of the four regimes in low-pressure wet steam.

Jump conditions across all types of aerodynamic waves are derived by Guha (1992a, 1994a). Figure 11 shows the predictions of the integral analysis compared with numerical solutions of the wave profile as discussed in §3.3. The dotted lines in figure 11 represent the jump conditions. The solid lines represent numerical solutions for three
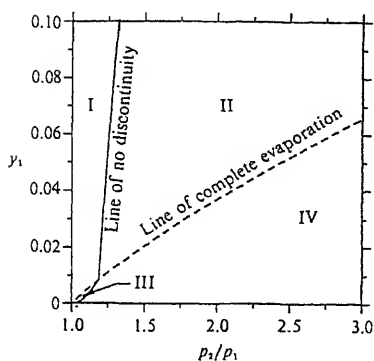
Figure 10. Phase diagram of different shock structures in low-pressure steam.

categories of shock waves: (I), partly dispersed with complete evaporation ($p_1 = 0.35$ bar, $M_{f1} = 1.6$, $y_1 = 0.05$, $r_1 = 0.1 \mu m$); (II) equilibrium partly dispersed ($p_1 = 0.35$ bar, $M_{f1} = 1.2$, $y_1 = 0.05$, $r_1 = 0.1 \mu m$); (III) equilibrium fully dispersed ($p_1 = 0.35$ bar, $M_{f1} = 0.97$, $y_1 = 0.1$, $r_1 = 0.1 \mu m$). Results from integral analyses agree remarkably well with solutions of the differential equations of motion, thereby confirming independent theoretical consistency.

It is interesting to study the various mechanisms of entropy creation inside a dispersed shock wave (Guha 1992a). In addition to the effects of viscosity and thermal conduction, entropy is created due to the relaxation processes. Figure 12 shows the rise in entropy across shock waves as a function of upstream frozen Mach number. Two cases are considered: dry steam, and, wet steam with upstream wetness fraction 0.1. Shock waves may occur in wet steam even when $M_{f1} < 1$, due to the existence of fully dispersed waves. In a partly dispersed wave, entropy rises across the frozen shock as well within the relaxation zone that follows. The figure shows that the contribution of the relaxation processes is very significant in the overall creation of entropy.

In an ideal gas, the total rise in entropy across a shock wave is *fixed* by Rankine–Hugoniot relations, the magnitudes of viscosity and thermal conductivity merely determine the thickness of the shock wave. It has been shown (Guha 1992a) that the total rise in entropy across a dispersed shock wave in vapour–droplet mixtures is similarly fixed by the jump conditions. In exact analogy with the role of viscosity, various relaxation processes and
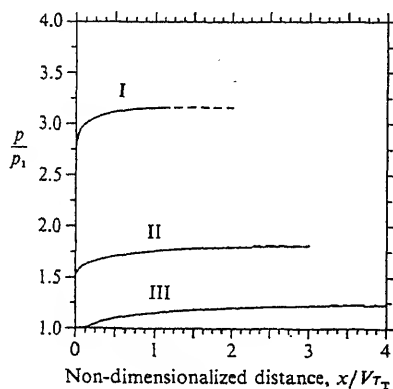
Figure 11. Comparison of asymptotic pressure profiles with generalised Rankine–Hugoniot equations.
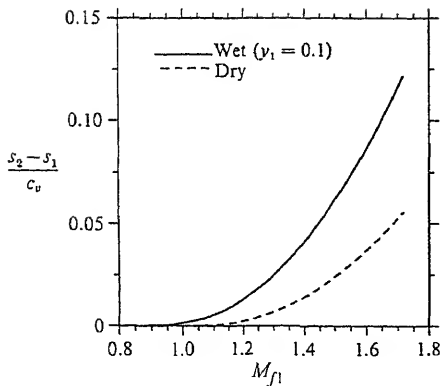
Figure 12. Entropy rise across shock waves in low-pressure steam.

their timescales determine the thickness of the shock wave. Overall jump conditions in any property, including entropy, can be determined without any explicit reference to the processes which make the 'jump' happen!

An integral analysis as well provide much insight into the unsteady development of shock waves. Details can be found in (Guha 1995; Guha 1992c).

3.6 *Interpretation of total pressure and temperature in two-phase flow*

In this section we discuss briefly some interesting effects of the non-equilibrium, interphase transfer mechanisms in a stagnation process in two-phase flow. Pitot measurements are often used for inferring velocity or loss (entropy generation) in multiphase mixtures. In single-phase fluids, the fluid is assumed to be brought to rest at the mouth of the Pitot tube *isentropically*. Hence flow Mach number and entropy generation (in steady, adiabatic flow) are uniquely determined by the total pressure measured by a Pitot tube, together with an independent measurement of the static pressure. (In supersonic flow in an ideal gas, application of Rankine–Hugoniot equations across the detached shock wave in front of a Pitot tube retains the utility of Pitot measurements for deducing flow Mach number and entropy generation.) Pitot measurements in a multiphase mixture, however, require careful considerations. Similar considerations are also needed for interpreting total temperature. Guha (1997a) gives the details of the physical considerations required and the description of a unified theory for the interpretation of total pressure as well as total temperature.

The solid particles or the liquid droplets respond to changes in temperature, velocity etc. of the gas phase through interphase exchanges of mass, momentum and energy. These are essentially rate processes and hence significant departures from equilibrium can take place if the rate of change of external conditions, imposed by the deceleration in the stagnating flow, is comparable to the internal time scales. Thus, for example, if the size of the liquid droplets or the solid particles is very small, then inertial and thermodynamic equilibrium between the two phases are maintained always, and a Pitot tube would measure the equilibrium total pressure, p_{oe}. On the other hand, if the size of the droplets or the particles is very large, all interphase transfer processes remain essentially frozen. The Pitot tube records the pressure which it would have recorded if the vapour phase alone was brought to rest from the same velocity. The total pressure in this case is termed the

frozen total pressure, p_{of}. Analytical expressions for p_{oe} and p_{of}, both in vapour-droplet and gas-particle flow, are given (Guha 1997a).

As an example, consider low-pressure wet steam with a typical wetness fraction of 10% and at a Mach number 1.5. Calculations show that $p_{of}/p = 3.3$ and $p_{oe}/p = 3.79$, where p is the static pressure. Therefore, in this particular example, the equilibrium total pressure is about 15% higher than the frozen total pressure.

It is expected that for intermediate sizes of the droplets or particles, the pressure recorded by the probe would neither be the equilibrium nor the frozen value. The imposed deceleration in front of the Pitot tube would cause the two-phase mixture to deviate from equilibrium conditions, both inertially and thermodynamically. *The deceleration process consequently ceases to be isentropic, as non-equilibrium exchanges of mass, momentum and energy between the two phases create entropy.*

Guha (1997a) has considered a large number of two-phase mixtures, both gas-particle and vapour-droplet, at subsonic as well as supersonic velocities for many different sizes of the droplets (or particles). In the supersonic case a detached frozen shock wave stands in front of the Pitot tube. The relaxation mechanisms in a gas-particle mixture are different from those in a vapour-droplet flow. Despite all these complexities and differences, it was possible with proper non-dimensionalization of flow parameters to adopt a *universal* plot, within acceptable tolerance, of non-dimensional total pressure, R_p, versus Stokes number, St (which is a non-dimensional representation of particle size). R_p and St are defined by, $R_p \equiv (p_o - p_{of})/(p_{oe} - p_{of})$ and $\text{St} \equiv \tau_I V_\infty / L$, where, V_∞ is the unperturbed velocity of the two-phase mixture far upstream of the measuring device, p_o is the pressure attained at the measuring point under non-equilibrium conditions (the total pressure which is measured) and L is a characteristic length (in subsonic flow L is related to the Pitot diameter, in supersonic flow L is related to the distance between the frozen shock wave and the Pitot mouth). Larger droplets or particles correspond to higher St.

Figure 13 shows the variation of R_p with St, which may be adopted as the Pitot correction curve usable at a wide range of subsonic and supersonic Mach numbers and for any two-phase mixtures (vapour-droplet or gas-particle). The variation is monotonic. *It should be noted that the denominator in the expression for R_p is calculated using the equilibrium thermodynamics, whereas the numerator is calculated using non-equilibrium equations.*
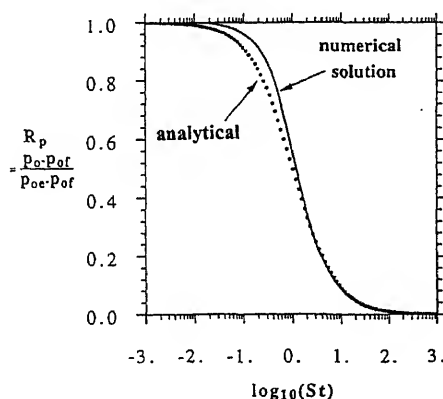
Figure 13. Near-universal plot of R_p versus St: comparison of a simple analytical theory with numerical solutions.
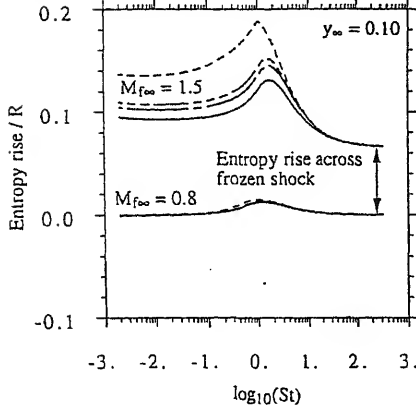
Figure 14. Entropy rise versus Stokes number in air with different solid particles $\delta = 0.1$ (——), 0.8 (— — —), 1.2 (– – –), 4.0 (——).

That the value of R_p, shown in figure 13, indeed tends to unity and zero in the appropriate limits of St, *demonstrates independent theoretical consistency of the calculation schemes.*

In addition to these numerical calculations, an *analytical* theory for determining total pressure under non-equilibrium conditions has been formulated. The analytical theory is simple and is amenable to direct physical interpretation. The theory shows that $R_p = 1/(1 + St)$. The predicted total pressure correctly reduces to the frozen total pressure in the limit of large Stokes number (large particles) and to the equilibrium total pressure in the limit of small Stokes number (small particles). Maximum dependence of the total pressure on Stokes number is observed when the Stokes number is of the order unity. The analytical result is also plotted in figure 13 for comparison. Under non-equilibrium conditions for intermediate St, the prediction of this equation compares very well with results from full numerical solution of the gas dynamic equations for two-phase mixtures.

Figure 14 plots the rise in mixture entropy, as mixtures of air and solid particles are decelerated by a measuring probe from their far upstream velocity to rest. Four different solid particles (hypothetical) with $\delta \equiv c_l/c_{pg} = 0.1, 0.8, 1.2$ and 4 are considered and the calculations are done for two Mach numbers. For the subsonic case ($M_{f\infty} = 0.8$), figure 14 shows that the rise in entropy is indeed maximum when St ~ 1, and is almost zero in the frozen and equilibrium limits. (Recall from figure 13 that the total pressures are different in these limits.) At $M_{f\infty} = 1.5$, the entropy rise is again maximum close to St ~ 1, but it has a finite value both at St $\rightarrow 0$ and at St $\rightarrow \infty$. The rise in entropy in the limit St $\rightarrow \infty$ is simply that across the frozen shock. (Since the same frozen shock is involved in all cases because the same $M_{f\infty}$ is used, this increase in entropy is the same for all four mixtures considered.) The rise in entropy in the limit St $\rightarrow 0$ is, however, different for different mixtures (it depends on the isentropic index of the *mixture* and hence on δ). However, it is shown (Guha 1992a) that if the particles come to equilibrium downstream of a frozen shock wave, then the entropy rise (across the shock plus the relaxation zone) is *not* dependent on the particle size (and hence on the relaxation times) but is determined completely by Rankine–Hugoniot equations for two-phase flow. This fact is reflected in the straight, horizontal portions of the curves (at $M_{f\infty} = 1.5$) in figure 14 in the limit St $\rightarrow 0$.

The rate of entropy production in a multiphase mixture is maximum when the Stokes number is of the order unity (in accordance with other results of relaxation gas dynamics), and a reduction in measured total pressure is not unequivocally related to a rise in entropy (as it is in steady, adiabatic flow of *single-phase* fluids).The fact that the total pressure decreases monotonically from p_{oe} to p_{of} as St changes form 0 to ∞ whereas the entropy rise is zero at both limits and has a maxima when St $\sim$ 1, demands care while interpreting Pitot measurements in multiphase flow.

4. Application of the non-equilibrium theory to steam turbines

4.1 *The formation of the liquid phase*

An introduction to the flow through steam turbines may be found in Guha (1995). In a multistage steam turbine used in power plants for generating electricity, the steam enters the low-pressure (LP) turbine cylinders as a dry superheated vapour but becomes wet towards the last stages. Experiments show that the water in turbines exists in two quite different forms. Usually more than 90% of the mass is concentrated in the fog, which consists of a very large number of very fine droplets (diameter 0.05–2.0 μm). The rest is in the form of coarse droplets which are very much larger (diameter 20–200 μm). Coarse water is formed as a small proportion of the fog (typically 2–3% per blade row) is deposited on the blade surface either due to inertial impaction or through turbulent diffusion. The deposited water is drawn towards the trailing edge by the steam flow (or centrifuged towards the casing on moving blades), where it is re-entrained in the form of large droplets. The large droplets cause blade erosion, but their thermodynamic and mechanical effects on the steam flow can nearly always be neglected.

Formation of new droplets occurs only over a small part of the expansion in a turbine. As a fluid particle flows through the machine, typically the complete nucleation process takes only 10–20 μs, as compared with a typical flow transit time of 5–10 ms through an LP turbine. Thus most of the expansion in the turbine simply involves condensation on existing droplets. However, nucleation is of crucial importance as it is the process which establishes the final fog droplet size distribution which, in turn, determines the subsequent departure from equilibrium affecting the flow behaviour, the magnitude of the wetness loss and the rate of fog droplet deposition on the blading forming coarse water. Once the droplet size distribution can be accurately predicted, the analysis of the wet steam flow through the rest of the turbine rests on more solid foundation. However, currently no theory exists (see § 4.2 for a novel theory) which gives even remote agreement with the available experimental measurements of the size distribution of fog droplets in turbines.

Our inability to understand the nucleation process in steam turbines is surprising given the success with which spontaneous condensation in laboratory nozzles and *stationary*, two-dimensional, laboratory cascade of steam turbine blades can be predicted using a synthesis of the classical theories of homogeneous nucleation and droplet growth with the conservation equations of gas dynamics. Such calculations have now been refined to the extent that the theory, amended by only a modicum of empiricism, gives acceptable
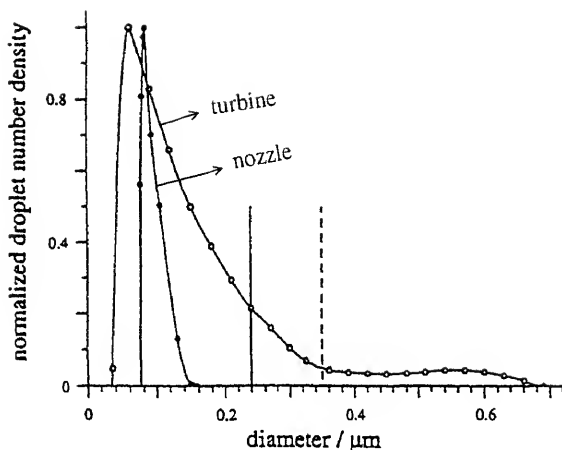
Figure 15. Typical droplet size distributions in a supersonic condensing steam nozzle (computed) and in the exhaust of a low pressure, electricity generating steam turbine (inferred from optical transmission data). • nozzle; ○ turbine ; – – – Sauter mean; —— mass mean. (Turbine measurements courtesy of PT Walters, National Power Technology and Environment Centre, Leatherhead, UK.)

agreement in terms of pressure distribution and mean droplet diameter with most experiments reported in the literature.

As shown in figure 15, calculations of droplet size spectra in condensing steam nozzles usually indicate a narrow distribution with a comparatively small mean droplet diameter strongly dependent on the local expansion rate near the Wilson point. Because of the narrowness of the distribution, optical experimental techniques based on measuring the attenuation of light of different wavelengths are unable to resolve the details of the spectrum but register results consistent with a near monodispersed droplet population. In turbines, however, experimental determinations of droplet size spectra give quite different results, the optical characteristics of the medium invariably indicating a broad, strongly skewed, distribution with a much larger mean diameter, typically in the range 0.2–0.6 μm. Quite often, the distribution is bimodal with a significant proportion of the total mass of liquid contained in a secondary population of droplets having diameters in the range 0.4–1.0 μm. A typical distribution, inferred from light attenuation measurements in a low-pressure steam turbine used for electricity generation, is also shown in figure 15.

Early attempts to describe the process of phase-transition in steam turbines followed the work of Gyarmathy who modelled the turbine by a series of one-dimensional nozzles. Reversion to equilibrium from the supersaturated vapour state was predicted to occur at a well-defined Wilson point in a particular blade passage giving rise to a near-monodispersed population of usually rather small droplets. The nucleation zone was thought to extend over a very short distance in the flow direction and calculations indicated that the mean droplet size should be very sensitive to turbine inlet conditions, small changes of which had the potential for displacing the Wilson point to new locations of quite different expansion rate. Later calculations of nucleation and droplet growth in two-dimensional turbine cascades displayed characteristics essentially similar to those found in one-dimensional condensation studies, albeit in more complicated and realistic fluid flow fields.

As noted above, optical measurements of wetness fraction and droplet spectra in turbines tell quite a different story. The measured spectrum is always broader, and the mean diameter larger, than conventional calculations indicate. Furthermore, the spectrum is comparatively insensitive to small changes in turbine inlet conditions: Measurements taken on the same machine over a period of years show excellent reproducibility. Finally, it appears

laboratory nozzles but generally occupies much greater distances in the flow direction. This behaviour can be deduced from optical measurements which show that the wetness fractions at certain locations in some turbines are considerably lower than the local equilibrium values.

It is therefore evident that the nucleation of water droplets in turbines involves phenomena which are not reproduced by laboratory experiments on nozzles and *stationary* cascades but nevertheless play a dominating role in the process of phase-transition in real machines. Possible explanations include nucleation in blade-wake vortices, heterogeneous condensation due to the effects of impurities in the steam and the effects of blade-row interaction unsteadiness. In § 4.2, consideration is given to the third of these possibilities, namely the effect of the unsteadiness due to the interaction of blade-wakes with downstream blade rows. It will be shown that a direct result of these interactions is to dramatically broaden the droplet size distribution giving a general shape and mean diameter much more in keeping with experimental measurements in real turbines. Furthermore, the theory predicts that the formation of the liquid-phase takes place in an essentially unsteady manner and may encompass a region in the flow comprising one or more complete turbine stages in a multi-stage machine. The theory presents a radically different perspective of nucleation in turbines from the generally accepted view and, if correct, should have a major influence on the future development of calculation procedures for non-equilibrium steam flows in turbines.

4.2 *Effects of unsteadiness on the homogeneous nucleation of water droplets in multi-stage steam turbines*

The details of the theory are given by Guha & Young (1994) and Guha (1995). The essence of the theory is that large-scale temperature fluctuations caused by the segmentation of blade wakes by successive blade rows have a dominating influence on nucleation and droplet growth in turbines. The fundamental premise is that, in passing through a multi-stage turbine, different fluid particles undergo different fluid mechanical experiences depending on the exact details of their passage through the machine and hence arrive at a given axial location with a wide variety of thermodynamic conditions. It is further assumed that, downstream of any turbine stage, the pressure of all the fluid particles would be near-uniform but their specific entropies (and hence static temperatures) would vary greatly depending on the dissipation experienced by a particular particle due to its being entrained in one or more blade boundary layers or lossy regions of the flow. However, although the path taken by a fluid particle is assumed to be random, the time-averaged dissipation of all the particles should agree with the overall loss distribution in the turbine. This is assumed known, either by direct measurement or from empirical loss correlations.

Figure 16 is a schematic diagram of the way in which the wakes from one blade row interact with, and are segmented by, the following row. It can be clearly seen that dissipation occurring in successive blade rows can become superposed in certain fluid particles (the darkly shaded areas).

A Lagrangian frame of reference is adopted and attention is focussed on a large number of individual fluid particles during their passage through the turbine. Homogeneous
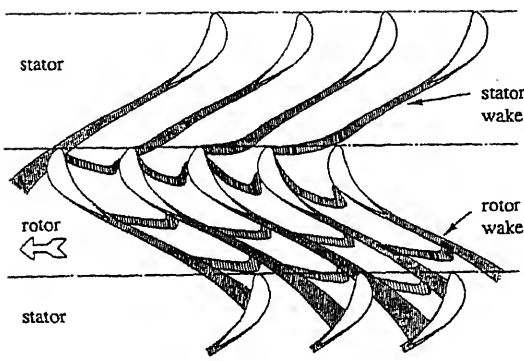
Figure 16. Schematic diagram of the passage of blade wakes through successive blade rows.

nucleation and growth of droplets in each fluid particle is assumed to be governed by classical theories. All fluid particles are assumed to experience the same pressure variation but those particles passing close to the blade surfaces suffer greater entropy production and therefore have higher static temperatures than those which pursue near-isentropic paths through the central portions of the blade passages. Particles which suffer high loss therefore nucleate later in the turbine than those which experience little dissipation. Condensation is thus viewed as an essentially random and unsteady phenomenon as the dissipation experienced by a fluid particle in one blade-row is assumed to be uncorrelated with its previous history. On a time-averaged basis, the condensation zone is spread over a much greater distance in the flow direction than a simple steady-flow analysis would indicate and may encompass several blade-rows depending on the number of stages in the machine. Predicted droplet size spectra show broad, highly-skewed distributions with large mean diameters and sometimes slight bimodality. These are all characteristics of experimentally measured spectra in real turbines. Conventional, steady-flow calculation methods, which predict a fixed Wilson point in a specific blade-row and a near-monodispersed droplet population, cannot reproduce any of these characteristics.

As shown in figure 17, a "loss profile" is constructed to represent the pitchwise distribution (from the suction to the pressure surface) of the loss in the blade-row. The pitchwise
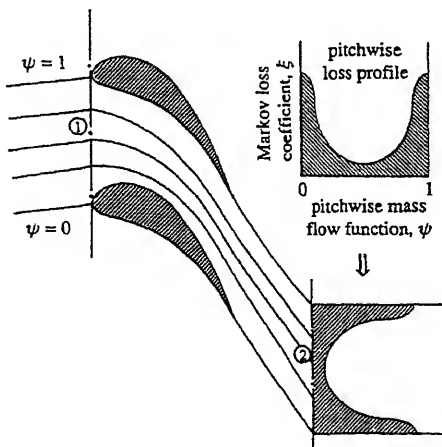
Figure 17. Specification of the pitchwise loss profile for a blade row.

loss profiles represent the time-averaged entropy increase along particular pathlines but individual fluid particles associated with the passage of wakes may exit from the blade-row at different conditions because, on entry, their static temperatures and velocities deviate from the mean.

For the results presented below, a single (circumferentially averaged) pressure-time variation based on an axisymmetric streamline-curvature throughflow solution was adopted for all fluid particles. Fluid particles are then launched at the turbine inlet (where the steam is superheated), all at the same stagnation temperature and pressure. At the entry to each blade row a random number is generated that specifies the pathline to be followed by the particle. The pathline, in turn, fixes the value of the polytropic efficiency (figure 17). The "black-box" (§ 2.2) is then applied along the pathline to calculate the subcooling and the droplet size distribution (if nucleation has taken place) at the downstream of the particular blade row. A new random number is then generated that specifies the pathline in the next blade row and the procedure is repeated. The "black-box" can deal with successive nucleations after the primary as a matter of course should the expansion be sufficiently rapid to generate the high levels of subcooling required.

For each fluid particle, the subcooling and droplet size distribution at all points of interest are recorded. It is then a straightforward matter to compute the time-mean wetness fraction and other statistical properties in order to obtain a quantitative picture of the process of phase-change and liquid growth throughout the machine. In a six-stage low-pressure turbine, some 10^4 fluid particle calculations are undertaken on *each* streamsurface to obtain converged statistical properties.

It has been assumed that the classical theories of homogeneous nucleation and droplet growth realistically describe the process of phase-change for *individual fluid particles*. In other words, if the pressure-time and temperature-time variations of a fluid particle during its passage through the turbine can be accurately specified, then it is assumed that the theories of nucleation and droplet growth correctly describe the rate of formation of the liquid-phase within that particle. For different fluid particles, phase-change will therefore be initiated at different locations in the machine, depending on the dissipation suffered by the particular particle. This is in contrast to the usual model of nucleation in turbines which assumes phase–change to be governed by the time-averaged fluid properties and therefore to occur at a single, well-defined location.

As an example calculation, the flow through the low pressure stages of a 320 MW turbine was analysed. The turbine was manufactured by the Italian company "Ansaldo" and the complete geometry is available in the literature (Guha & Young 1994; Guha 1995). The LP section has six-stages. Each stage consists of a stator followed by a rotor and hence there are twelve blade-rows altogether in the turbine.

The physical characteristics are best explained by adopting a Lagrangian viewpoint of a fluid particle as it passes through the turbine. As described in the previous sections, different fluid particles experience different amounts of dissipation and heat transfer, depending on the particular pathline followed. Two limiting cases can be identified. At one extreme are the fluid particles which always follow the mid-pitch pathline in *each* blade-row and consequently suffer no dissipation. They pursue an isentropic path to the Wilson point. At the other extreme are those particles which negotiate the regions of maximum loss in *each*
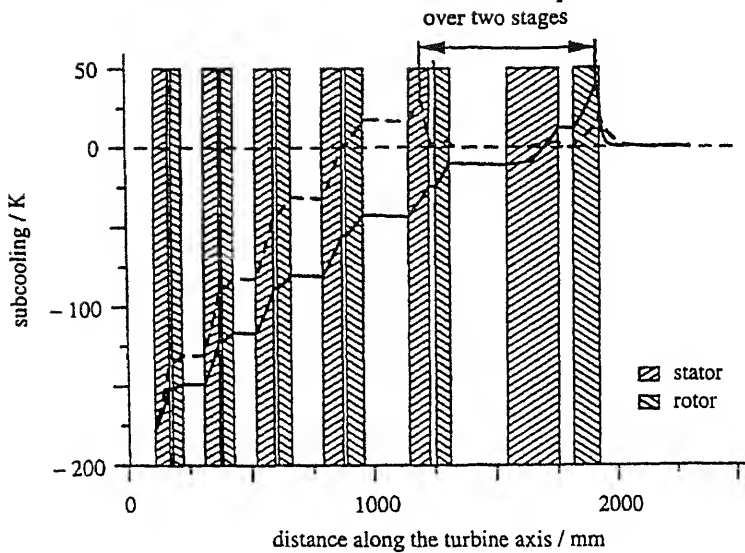
Figure 18. Axial variation in the Ansaldo turbine of vapour subcooling for two fluid particles representing the extreme cases of zero dissipation (– – –) and maximum dissipation (—).

blade-row. Other fluid particles experience levels of loss intermediate between these two extremes.

Figure 18 shows the calculated variation of the vapour subcooling ΔT associated with the two extreme cases of zero and maximum dissipation. The subcooling of the fluid particles change as a result of the competing effects of the three physical processes: expansion, condensation and dissipation. Consider first the case of the fluid particle suffering no dissipation. Here, one of the mechanisms for altering the subcooling, i.e., dissipation, is absent. The fluid particle is superheated (negative ΔT) at the turbine inlet. Its subcooling increases in each blade-row due to expansion but remains almost constant between the rows. It attains the Wilson point in the stator of the fifth stage and subsequently experiences an exponential decrease in ΔT due to the extremely rapid liberation of latent heat. ΔT increases significantly again in the last rotor where the expansion rate is too high to be offset by the counteracting effect of condensation. Much the same history is repeated for the fluid particle experiencing the maximum dissipation. Here, however, dissipation opposes the increase of the subcooling throughout the flow field. Consequently, the Wilson point occurs much further downstream (in the rotor of the last stage). Other fluid particles, experiencing intermediate amounts of dissipation, attain their Wilson points at intermediate locations between the two extremes. The region of nucleation thus covers (in a randomly unsteady manner) almost two complete turbine stages as opposed to being restricted to a very narrow zone in a specific blade-row.

A reliable "rule of thumb" is that *Wilson points occurring at locations of higher expansion rate result in smaller droplets*. The variable location of the Wilson point therefore results in large variations in *mean* droplet diameter. It should be understood, however, that the mean droplet diameters in the two limiting cases *do not necessarily represent the*
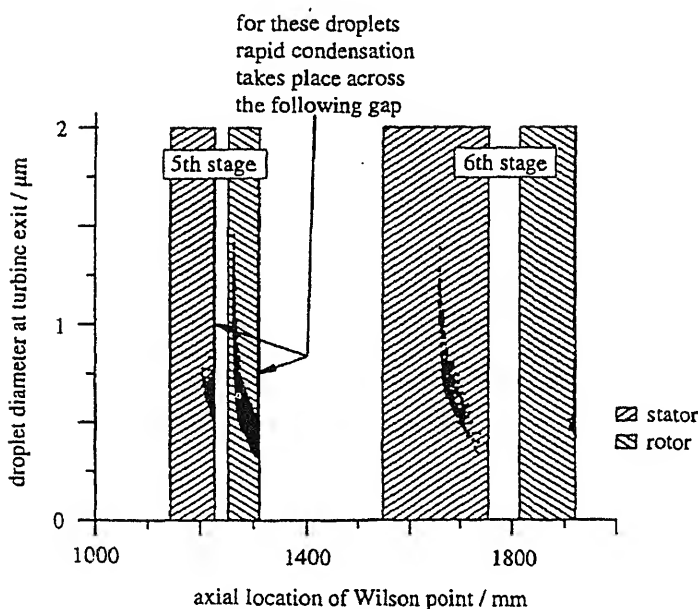
Figure 19. Computed Sauter mean droplet diameter at Ansaldo turbine outlet as a function of the axial location of the Wilson point for all fluid particles.

extreme limits of droplet size produced in the machine. The droplet size is dependent on the local expansion rate which does not vary monotonically with distance between the extreme locations of the Wilson points.

Figure 19 shows the location of the Wilson points for the totality of particles considered (10^4). Each point on this diagram corresponds to an individual fluid particle. The abscissa denotes the axial location of the Wilson point and the ordinate denotes the *mean diameter* of droplets within the fluid particle *on its arrival at the turbine outlet.* (It should be appreciated that each fluid particle is actually composed of a polydispersed distribution of droplet sizes, details of which, although retained by the computational procedure, cannot be included in a single diagram such as figure 19.) It can be seen that the majority of fluid particles nucleate either in the fifth-stage rotor or in the sixth-stage stator. (The absence of Wilson points in the first part of the sixth-stage stator results from the very low expansion rate there, see (Guha & Young 1994; Guha 1995).) In each blade-row, the mean diameter of the droplets becomes progressively smaller as the Wilson point moves towards the trailing-edge, as the rate of expansion tends to increase monotonically to each blade throat. However, interesting behaviour results for those fluid particles that reach the trailing-edge plane with subcoolings and nucleation rates which, although moderately high, are still insufficient to cause complete reversion to equilibrium. For these particles, the trailing-edge marks the Wilson point (i.e., the cessation of nucleation) even if comparatively few droplets have yet been produced. Reversion to equilibrium then occurs within the following gap by condensational growth on existing droplets. Because there is plenty of time available and because the droplet number density is low, these droplets may grow to very large sizes as shown in figure 19. Such fluid particles are then prime candidates for secondary nucleations in succeeding blade-rows because their liquid surface area is insufficient to
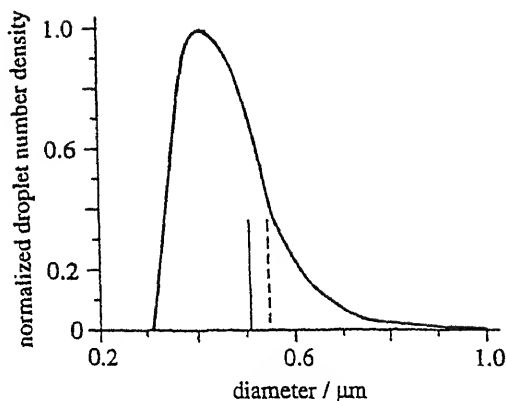
Figure 20. Computed time-averaged droplet size distribution at outlet of the Ansaldo turbine. —— mass mean; - - - Sauter mean.

offset (by condensation) the opposing effect of increased subcooling due to the rapid expansion.

An imaginary probe with unlimited resolution in space and time, sited at the turbine outlet, would register the complete droplet size distribution for each fluid particle it encounters. Real probes based on the measurement of attenuated or scattered light, however, record only sufficient information to deduce, at most, the time-averaged droplet size distribution (and sometimes only the time-averaged Sauter mean diameter). In order to compare the theoretical predictions with such measurements, a theoretical time-averaged droplet size distribution may be constructed at any axial location in the turbine if the diameters and number density of droplets in all the 10^4 fluid particles considered are recorded by the computer for subsequent processing. The calculated time-averaged droplet size distribution for the Ansaldo turbine at outlet is shown in figure 20. The spectrum is polydispersed and highly-skewed (i.e., there is a large difference between the mean and most probable diameters) and resembles the shape of similar spectra measured in real turbines (figure 15). This is very significant, as no existing steady-flow calculation procedure can predict such a high degree of polydispersion.

Unfortunately, no measurement of the droplet size distribution is available for the Ansaldo turbine, although the time-averaged Sauter mean diameter of the droplets has been measured. The measured Sauter mean diameter at mid-span is about 0.4 μm, which is a little smaller than the calculated value of 0.55 μm, shown in figure 20. However, allowing for the uncertainties and approximations in the calculation scheme, the level of agreement is extremely encouraging. Of course, many more experimental comparisons are required before it is possible to assert conclusively that the important physical processes are being successfully modelled by the theory presented.

In conclusion to this section, a theory has been developed for predicting the effect of temperature fluctuations on the homogeneous nucleation and growth of water droplets in multi-stage steam turbines. The fluctuations result from the segmentation of blade-wakes by successive blade-rows and the amplitude of the fluctuations increases with the number of stages. According to the model, the mechanics of nucleation in multi-stage turbines are quite different from the predictions of conventional steady-state theories of phase-change. For example, the nucleation zone may encompass (in a randomly unsteady manner) several blade-rows (as opposed to being isolated at a particular position in a specific blade-row).

The inherent unsteadiness of the process also results in a highly-skewed, polydispersed (sometimes bimodal) time-averaged droplet size distribution, having similar characteristics to spectra measured in real turbines. The next step would be to include, in the calculation scheme, the effects of circumferential variation in pressure within the blade passages. More details may be found in Guha & Young (1994) and Guha (1995).

The well-known books on multi-phase flows (e.g. by N A Fuchs and by G B Wallis) describe the long, inexhaustible list of multi-phase phenomena demonstrating their all-pervasive occurrence. Here, we have considered only a few important topics, the selection inevitably being biased by the author's own interests. (See Guha 1997b for a new theory for particle transport in turbulent flow.) We have used a complementary combination of analytical and computational techniques, and differential and integral treatments in order to model fundamental processes occurring in two-phase mixtures as well as to explain observed phenomena and experimental findings. This is an exciting, rewarding and potent field – so many interesting and important things remain to be done!

References

Guha A 1992a Jump conditions across normal shock waves in pure vapour-droplet flows. *J. Fluid Mech.* 241: 349–369

Guha A 1992b Structure of partly dispersed normal shock waves in vapour-droplet flows. *Phys. Fluids* A 4: 1566–1578

Guha A 1992c The physics of relaxation processes and of stationary and non-stationary shock waves in vapour-droplet flows. In *Transport phenomena in heat and mass transfer* (ed.) J Reizes (Amsterdam: Elsevier) pp 1404–1417

Guha A 1994a A unified theory of aerodynamic and condensation shock waves in vapour-droplet flows with or without a carrier gas. *Phys. Fluids* 6: 1893–1913

Guha A 1994b Thermal choking due to nonequilibrium condensation. *Trans. ASME J. Fluids Engg.* 116: 599–604

Guha A 1995 *Two-phase flows with phase transition.* VKI Lecture Series 1995-06 (von Karman Institute for Fluid Dynamics) pp 1–110

Guha A 1997a A unified theory for the interpretation of measured total pressure and total temperature in multiphase flows at subsonic and supersonic speeds. *Proc. R. Soc. London* 453: (in press)

Guha A 1997b A unified Eulerian theory of turbulent deposition to smooth and rough surfaces. *J. Aerosol Sci.* 28(8): (in press)

Guha A, Young J B 1989 Stationary and moving normal shock waves in wet steam. In *Adiabatic waves in liquid-vapour systems* (eds) G E A Meier, P A Thompson (Berlin: Springer) pp 159–170

Guha A, Young J B 1991 Time-marching prediction of unsteady condensation phenomena due to supercritical heat addition. *Proc. Conf. Turbomachinery: Latest Developments in a Changing Scene* (London: Institute of Mechanical Engineers) paper C423/057, pp 167–177

Guha A, Young J B 1994 The effect of flow unsteadiness on the homogeneous nucleation of water droplets in steam turbines. *Philos. Trans. R. Soc.* 349: 445–472

Young J B, Guha A 1991 Normal shock-wave structure in two-phase vapour-droplet flows. *J. Fluid Mech.* 228: 243–274

The vortex liquid piston engine and some other vortex technologies

M GOLDSHTIK, F HUSSAIN and R J YAO

Department of Mechanical Engineering, University of Houston, Houston, TX 77204-4792, USA
e-mail: FHussain@uh.edu

Abstract. By exploiting three unique characteristics of confined swirling incompressible flows – centrifugal acceleration, internal separation or recirculation zones near the axis, and *bistability* (i.e. rarefied and condensed stable states) of multi-phase flows – we developed several innovative *vortex machines* which will revolutionize mechanical technologies in a variety of industries. The machines utilizing these features include: *Vortex Engine*, *Vortex Thruster*, *Vortex Suction Device*, *Vortex Chemical Reactor*, *Bubbling Centrifuge* and *Vortex Mill*. As a specific example, we describe here in some detail the development of a liquid piston engine, including analysis of its hydrodynamic and thermodynamic features. We have designed a laboratory 'cold' model and performed detailed experimental, theoretical and numerical analyses to study the role of the controlling parameters and are now ready to test a 'hot' model. In addition, we mention a few other vortex technologies of interest to us.

Keywords. Multi-phase flows; incompressible flows; vortex machines; vortex liquid piston engine; swirling flows.

1. Introduction: Rotating single- and multi-phase flows

While the proposed applications are innovative, rotating flows within vortex chambers have been investigated extensively. As shown schematically in figure 1, the tangential entry of fluid into a typical cylindrical vortex chamber, via the guiding vanes around the periphery, causes swirling motion of the fluid (mixture). In the core region of the flow away from the walls, the radial pressure gradient balances the centrifugal force; this is the so-called *cyclostrophic balance*. Near the end walls, the centrifugal acceleration is very small due to the boundary layer effect but the radial pressure gradient is essentially the same as in the core region. Thus, the lack of cyclostrophic balance causes the pressure gradient to

---

A list of symbols used appears at the end of the paper.
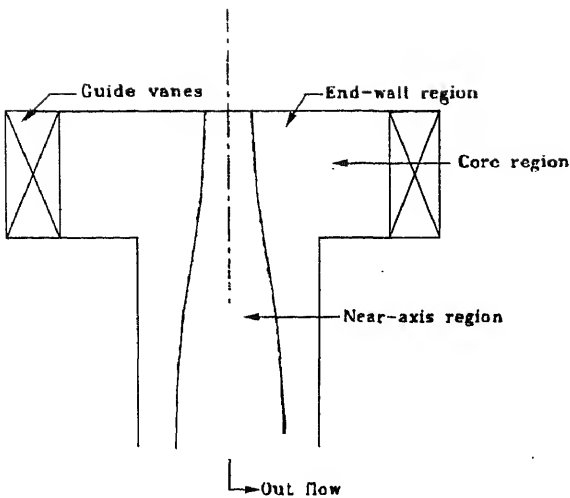
Figure 1. Schematic of various flow regions in an axisymmetric vortex chamber.

generate a secondary flow towards the axis along the end walls, balanced by the frictional force. This leads to an end-wall boundary layer with a substantial inflow radial speed v_r, the maximum of v_r (and of axial and azimuthal velocities) occurring inside the boundary layer. This inflow of the rotating fluid along the end wall prevents the typically desirable prolonged retention of small particles within the chamber, because they are carried away by the fluid flow along the axis. However, by appropriately profiling the end walls (we have developed an equation for the end wall profile using hydrodynamic analysis and laboratory tests), small particles can be retained within the chamber for longer durations. At practical operating speeds, the flow in the near-axis region is extremely complex, involving very high turbulence and noise-producing axial oscillations. A detailed study of the flow patterns in
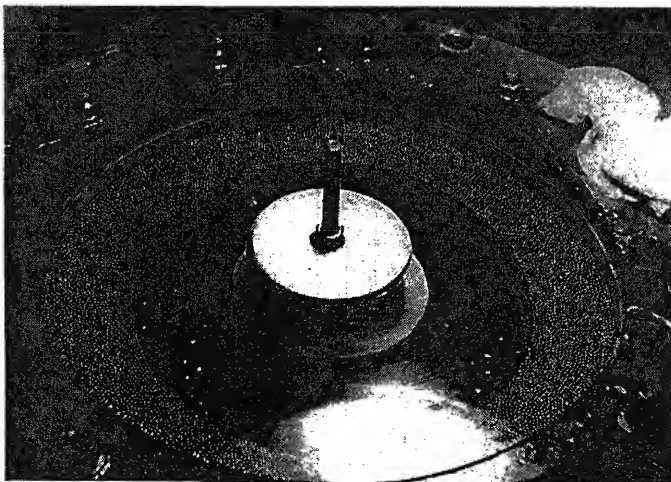
Figure 2. Condensed stable state of a mixture of liquid-solid particle in a vortex chamber. Note the dense, homogeneous particle layer near the periphery, forming a rotating axisymmetric fluidized bed.
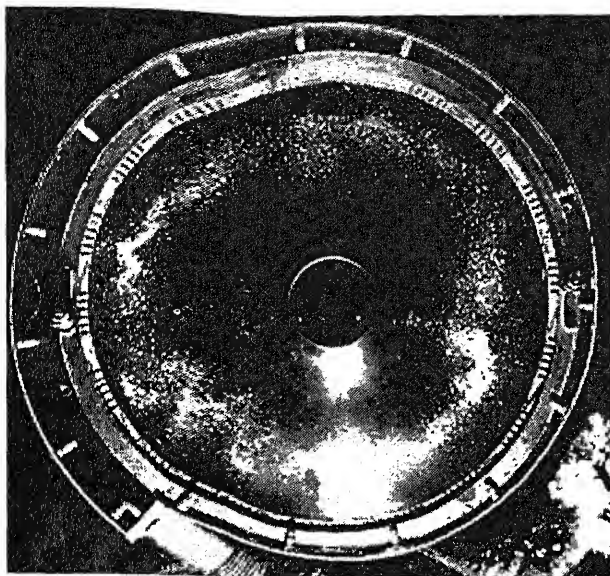
Figure 3. Rarefied stable state with small particles rotating in circular trajectories in a vortex chamber.

this and other regions of the vortex chamber for various operating parameters is necessary to develop more efficient designs.

Experimental and theoretical studies (Batchelor 1967; Greenspan 1969; Goldshtik 1981, 1984; Gupta *et al* 1984; Hussain 1986) show that rotating flows possess three fundamental features, described below.

High centrifugal force: Typical inlet gas velocities (≈ 100 m/s) into a chamber of radius ≈ 0.1 m produce a centrifugal acceleration $a_c \approx 10^4$ g. This high centrifugal force is central to all applications discussed here, and ensures stability of the central cavity and motion of solid particles/fluid bubbles in circular trajectories inside the cylindrical vortex chamber.

Near-axis flow: Near the axis of rotation, if the speed is high enough, a gaseous cavity is formed when a liquid is used as the working fluid, and a recirculation zone (such as vortex breakdown bubble or internal separation, i.e. separation away from any wall), forms if the fluid is a gas. The complex flow pattern in this region depends on the operating conditions (such as Reynolds number and Rossby number) and the chamber end-wall profile, and considerably influences the flow in the rest of the chamber. A clear understanding of the flow in this region based on rigorous hydrodynamic analysis is essential for optimal design of these machines, particularly the vortex engine and vortex thruster.
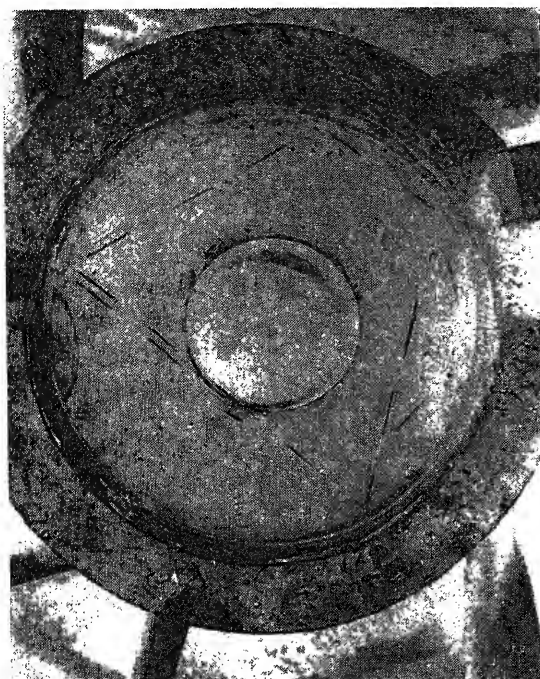
Figure 4. Rarefied stable state with large particles, which move in polygonal paths and collide with the outer wall, forming the basis for a vortex mill.

Condensed stable state: At high particle concentrations, the denser medium (solid or liquid) forms a tightly packed layer near the periphery of the vortex chamber and provides a very large interface area between the two media; figure 2 illustrates a condensed particle layer.

Rarefied stable state: At lower concentrations, the particles are distributed homogeneously and move in circular orbits in the lighter fluid medium (figures 3, 4). The radii of the orbits depend on the particle size. For solids, larger particles collide with the chamber walls and are ground to smaller sizes (figure 4); this leads to the proposed vortex mill (see §8).

In this paper we focus entirely on the liquid piston engine; the other machines will be considered very briefly at the end.

2.   Vortex engine

Conventional piston engines have relatively complicated designs and undergo greater wear and tear due to moving mechanical parts. An external combustion, flat-top liquid piston Stirling engine (West 1983) is very simple and does not employ complicated mechanical parts. The main drawback of such an engine is the instability of the liquid piston's top surface. When the piston is near its top dead centre, its speed becomes zero, but acceleration is maximum. Under these conditions, a stability criterion is that the ratio of gravitational
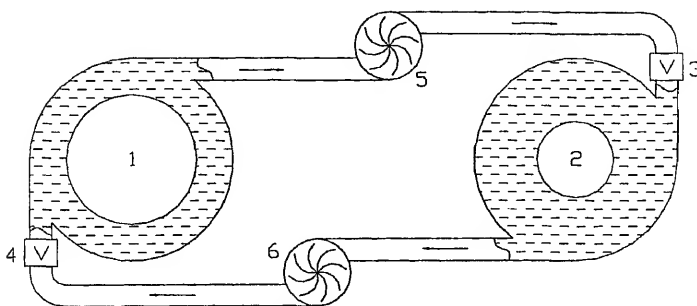
Figure 5. Working principle of the liquid piston engine (see Goldshtik 1992); 1, 2 – oscillating cavities; 3, 4 – check valves; 5, 6 – hydromotors.

acceleration to fluid acceleration must be greater than unity so that no liquid will leave the surface of the liquid piston. Therefore, the frequency of the engine must be less than one hertz in order to avoid the instability of the flat liquid piston surface. Accordingly, the efficiency of this engine is very low, about one percent, and the power of this engine is only several watts at best.

These limitations of the flat top liquid piston engine are overcome by employing a rotating liquid piston (Goldshtik 1992), where centrifugal acceleration is used instead of gravity. This is an internal combustion engine with two circular cylinders (i.e. vortex chambers) that are partially filled with a fixed volume of liquid and are connected tangentially by two channels containing hydromotors (see figure 5). Each of these circular cylinders has a top and a bottom cover, a system for intake of fuel and air, and an associated exhaust system. Each cylinder may have either an electric spark plug or may work in diesel mode (via a fuel spray injector). By using tangential entry, the liquid rotates at high speed within the cylinder and creates a vertical cylindrical gas cavity around the axis of the rotating liquid. Rotation is used to stabilize the liquid–gas interface of the cylindrical cavity, which functions like the piston top in a conventional internal combustion engine compressing the fuel–air mixture. This cavity is the combustion chamber into which the fuel–air mixture is injected. When the mixture is ignited, the increased pressure in the cavity forces some of the liquid out through a tangential channel, through a hydromotor, and then tangentially into the second cylinder wherein it keeps its swirling motion; this sequence is repeated in the second cylinder and the liquid then flows back into the first cylinder. In this manner, as a result of the pressure from combustion, the liquid is transferred back and forth between the two cylinders at a frequency that can be controlled by changing the system parameters. The unidirectionally rotating hydromotors extract energy for mechanical drive. In summary, this device works without any moving mechanical part, with the exception of the hydromotors and valves to control flow timing of liquid, fuel/air mixture, and products of combustion.

In this paper, we focus on the detailed theoretical description of the engine, and also present some preliminary numerical and experimental results. Results of the engine's numerical optimization and experimental investigations will be published in the

3.   Simple theoretical model of vortex liquid piston engine (VLPE)

3.1   *Principal dynamical equation*

The model assumes an unsteady flow of a liquid in two identical interconnected cylinders, with near-axis cylindrical cavities filled with an ideal gas and having time-dependent radii $r_1(t)$ and $r_2(t)$. The flow between the cylinders generates a self-sustained oscillation due to combustion-induced thermal expansion and contraction processes running in opposite phases. As a first step, viscosity is neglected; viscous effects are considered later in both theory and numerical simulations.

For axisymmetric flow, we have the following relationships for radial and tangential velocities:

$$v_r = Q/r \quad \text{and} \quad v_\varphi = \Gamma/r, \tag{1}$$

where $Q = Q(t)$ and $\Gamma = \text{const.}$ are related to the dimensional physical volume flow rate Q_v and circulation Γ_R by

$$Q_v = 2\pi h Q \quad \text{and} \quad \Gamma_R = 2\pi \Gamma, \tag{2}$$

where h is the chamber height. Conservation of the fluid volume gives the relation

$$r_1^2 + r_2^2 = 2\sigma R^2, \tag{3}$$

where R is the cylinder radius and σ is the total gas/cylinder volume ratio ($1 - \sigma$ is the total volume fraction of liquid). The radial velocity of the liquid–gas surface is $v_r = dr/dt$. From (1), we have

$$Q = r(dr/dt). \tag{4}$$

With the help of (3) and (4) we can obtain

$$Q_1 = -Q_2 = Q. \tag{5}$$

For an incompressible, inviscid liquid in unsteady, axisymmetric flow, we can use Euler's equation in cylindrical coordinates:

$$\frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} - \frac{v_\varphi^2}{r} = -\frac{1}{\rho}\frac{\partial p}{\partial r},$$

$$\frac{\partial v_\varphi}{\partial t} + v_r \frac{\partial v_\varphi}{\partial r} + \frac{v_r v_\varphi}{r} = 0,$$

$$\frac{1}{r}\frac{\partial(r v_r)}{\partial r} = 0. \tag{6}$$

The last two equations (6) are satisfied by (1) if Γ is constant. Substituting (1) into the first equation (6), and integrating from r to R, we obtain

$$\frac{dQ}{dt}\ln\frac{R}{r} - \frac{Q^2 + \Gamma^2}{2}\left(\frac{1}{r^2} - \frac{1}{R^2}\right) = \frac{p_r - p_R}{\rho}, \tag{7}$$

where p_r is the gas pressure in the cavity and $p_R = p(R)$ is the cylinder wall pressure.

After writing (7) for both cylinders, subtracting and taking into account (5), we get

$$\frac{dQ}{dt}\ln\frac{R^2}{r_1 r_2} - \frac{Q^2 + \Gamma^2}{2}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right) + \frac{p_{R_1} - p_{R_2}}{\rho} = \frac{p_{r_1} - p_{r_2}}{\rho}. \tag{8}$$

The pressures p_{R_1} and p_{R_2} are not equal; their difference is the available head for hydromotors. We characterize it approximately by the standard load coefficient ζ, which is assumed to be a constant, related to the pressure drop Δp across the hydromotor by

$$\Delta p = \tfrac{1}{2}\zeta(\rho/R^2)Q|Q|. \tag{9}$$

Now we choose the chamber radius R as a length scale, the tangential velocity at the chamber entrance $V = v_\varphi(R)$ (so that $\Gamma = RV$) as the velocity scale, and nondimensionalize the variables. The main nondimensional variable $y(t)$ is introduced by

$$r_1^2 = R^2(\sigma + y), \quad r_2^2 = R^2(\sigma - y), \quad \cdot \tag{10}$$

where y can be considered as a volumetric deviation from the equilibrium state ($y = 0$). Note that, because of (10), (3) is satisfied identically. The nondimensional time τ is

$$\tau = \frac{2V}{R}t, \quad \text{so} \quad \frac{d}{dt} = \frac{2V}{R}\frac{d}{d\tau}. \tag{11}$$

From (4) and (11)

$$Q = VR\dot{y}, \quad \text{and} \quad dQ/dt = 2V^2\ddot{y}, \tag{12}$$

where the dot means differentiation with respect to τ. Substituting (9)–(12) into (8), we obtain

$$\ddot{y}\ln\frac{1}{\sigma^2 - y^2} + \frac{\dot{y}^2 + 1}{\sigma^2 - y^2}y + \frac{\zeta}{2}|\dot{y}|\dot{y} = \Pi(f_1 - f_2), \tag{13}$$

where $f_1 - f_2 = (p_{r_1} - p_{r_2})/p_0$, $\Pi = p_0/\rho V^2$, and p_0 is the pressure scale, e.g. minimum pressure in the thermodynamical cycle used. Equation (13) is the principal dynamical equation for VLPE. It is similar to Rayleigh's well-known equation for bubble dynamics, but in contrast, (13) has no solutions for unbounded fluid; it does only in bounded cylindrical domain. Equation (13) is not a closed system because its right hand side is still unknown. Thermodynamic analysis is necessary to determine it.

3.2 Thermodynamics

From the first law of thermodynamics for ideal gases,

$$\delta q = c_v dT + p dV, \tag{14}$$

where δq is the heat supply, c_v is the specific heat at a constant volume, and T and V are temperature and volume respectively. Also for an ideal gas,

$$pV = RT, \tag{15}$$
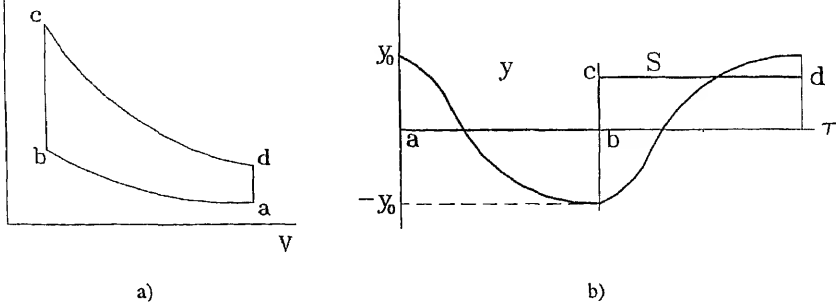
Figure 6. Otto cycle and time-process.

where R is the universal gas constant, i.e. $R = c_p - c_v = c_v(k - 1)$ and $k \equiv c_p/c_v$ is the adiabatic exponent.

Equation (13) has relative volume y and pressure f as dependent variables. Therefore, it is convenient to choose gas volume $V = \pi r_1^2 h = \pi R^2 h(\sigma + y)$ as the basic thermodynamic variable. Removing T from (14) by using (15) yields

$$\delta q = (V\,dp + kp\,dV)/(k - 1). \tag{16}$$

Assuming that the amount of heat δq is given, we can write a closed system equation from (14)–(16), in terms of entropy S (using $\delta q = T\,dS$ for a reversible process),

$$\frac{p}{p_0} = \left(\frac{V_0}{V}\right)^k \exp\left(\frac{k-1}{R}S\right), \tag{17}$$

where V_0 is the volume at the pressure $p = p_0$ and S_0 is taken to be zero for convenience.

For a cyclic working engine, the function $S(t)$ must be a periodic function of time. The specific form of $S(t)$ depends on the way heat is supplied, i.e. on the thermodynamic cycle. Most conventional engines operate on the Otto cycle, where heat addition begins during the phase when the piston is near its top dead centre (TDC). At this moment, the volume is near the minimum and combustion occurs very rapidly. In our vortex engine, this phase corresponds to the minimum cavity radius when the radial velocity is zero.

Let us consider, for example, an ideal Otto cycle, consisting of two adiabatic and two isochoric processes (figure 6a). Let this process begin from position a with the maximum volume, when $y = y_0$. The adiabatic compression takes place from a to b with $S = 0$. Then at the minimum volume (position b), the system suddenly absorbs heat and goes to position c. The entropy increases by a jump, after which there is an adiabatic expansion from c to d and then exhaust from d to a. The corresponding time-process is shown in figure 6b (the curve denotes the periodic condition (21), and the other cylinder is 90° out-of-phase with this one).

For the Otto cycle the value $f_1 - f_2$ from (13) can be found as

$$f_1 - f_2 = \left(\frac{\sigma + y_0}{\sigma + y}\right)^k - \left(\frac{\sigma + y_0}{\sigma - y}\right)^k \exp\left(\frac{k-1}{R}S\right), \tag{18}$$

where $S =$ const. during the adiabatic half-period. Note that σ and y_0 determine the compression ratio ϵ

$$\epsilon = (\sigma + y_0)/(\sigma - y_0). \tag{19}$$

The Otto cycle is chosen as an example only; real cycles are smoother. Equation (13), which governs the dynamics of our vortex liquid piston engine, has three parameters: σ (the total gas/cylinder volume ratio), ζ (the load coefficient), and Π (the pressure parameter). Two additional parameters, S (the heat parameter) and y_0 (the magnitude of oscillations), are contained in (18). Thus the system is characterized by five independent parameters, which are large for a complete parametric analysis. Some physical conditions will be used to reduce the number of independent parameters.

3.3 *Physical conditions*

Periodicity condition: Equations (13) and (18) form a closed autonomous system with the following initial conditions

$$y(0) = y_0 \quad \text{and} \quad \dot{y}(0) = 0, \tag{20}$$

where the initial time $\tau = 0$ is chosen at maximum $y(\tau) = y_0$. The value y_0 in (18) is not arbitrary but has to be determined from the periodic condition. We assume that in both cylinders, the processes are identical but only shifted by half a period $\frac{1}{2}\tau_c$, i.e.,

$$r_2(\tau + \tfrac{1}{2}\tau_c) = r_1(\tau).$$

Taking into account (10), we obtain

$$y(\tau + \tfrac{1}{2}\tau_c) = -y(\tau). \tag{21}$$

Then from (21), we get

$$y(\tfrac{1}{2}\tau_c) = -y_0 \quad \text{and} \quad \dot{y}(\tfrac{1}{2}\tau_c) = 0. \tag{22}$$

According to (21), (13) should be solved only in the interval $0 \le \tau \le \frac{1}{2}\tau_c$ but the value y_0 must satisfy (22); this can be achieved by an iterative procedure. It is the principal difference in comparison with the conventional internal combustion engine, where the piston amplitude and compression ratio are fixed.

Combustion condition: Let us consider an Otto cycle consisting of the two adiabatic curves (ab and cd) and two isochores (bc and da) shown in figure 6a. Heat is received along bc and is rejected along da. Along the process bc, $\delta q = c_v dT$ and $q = c_v(T_c - T_b)$; thus

$$q = q_f/(1 + n_0\alpha), \tag{23}$$

where q_f is enthalpy of reaction, α is the relative air/fuel ratio, and n_0 is the stoichiometric (or chemically correct or theoretical) air/fuel ratio. Thus $n_0\alpha$ gives the actual air/fuel ratio. Since $q = c_v(T_c - T_b)$ we have

$$T_c - T_b = q_f/[c_v(1 + n_0\alpha)] \tag{24}$$

On the other hand, for an isochoric process

$$S_c - S_b = c_v \ln (T_c/T_b) = c_v \ln E, \tag{25}$$

where

$$E = 1 + \frac{q_f}{c_v(n_0\alpha + 1)T_0\epsilon^{k-1}} = 1 + \frac{E_0}{\epsilon^{k-1}}. \tag{26}$$

For a typical hydrocarbon fuel, $q_f \cong 20{,}000$ Btu/lb $\cong 46{,}500$ kJ/kg. For typical cases using hydrocarbons, n_0 is about 15 kg air/kg fuel, $T_0 = 300$ K, and $\alpha = 1.25$. So $E_0 \cong 11$, and for an Otto cycle $E = 1 + 11\epsilon^{1-k}$. We can rewrite (18) in terms of E:

$$f_1 - f_2 = \left(\frac{\sigma + y_0}{\sigma + y}\right)^k - \left(\frac{\sigma + y_0}{\sigma - y}\right)^k E, \tag{27}$$

where E is given by (26), and (13) can be solved accordingly.

Kinematic condition: The flow in a vortex chamber is created by a special device, a *swirler*, which in principle can move or be at rest. In the case of a porous rotating cylinder, the flow in the vortex chamber is characterized by two independent functions $Q(t)$ and $\Gamma(t)$. Of course, in case of a fixed swirler, Q and Γ are not independent. This can be illustrated by a steady flow in a vortex chamber. Let the swirler contain a number of narrow tangential slits. The flow rate into the chamber can be approximated as $Q_v = nbhV$, and according to (2) we have

$$Q = (nb/2\pi)V, \tag{28}$$

where n is the number of slits, each of width b (figure 7).
    On the other hand, according to (1),
$$Q = R|v_r(R)|, \tag{29}$$
where $v_r(R)$ is the azimuthally averaged radial velocity. From (1) it follows that the design parameter,

$$\theta = \frac{Q}{\Gamma} = \frac{|v_r|}{v_\varphi} = \frac{nb}{2\pi R}, \tag{30}$$
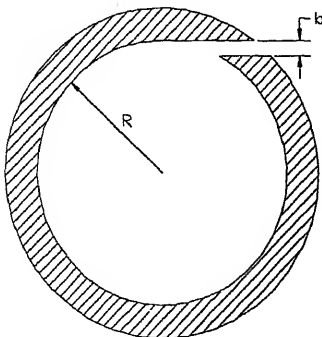
Figure 7.   Tangential slit.
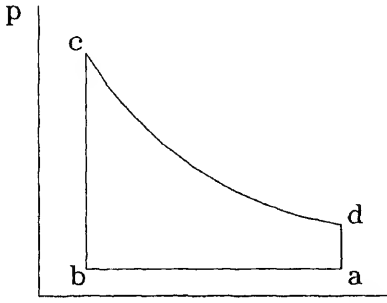
Figure 8. Semi-Otto cycle.

is a purely geometric characteristic of the vortex chamber which relates Q to Γ. Equation (30) shows that, once a design is chosen, θ should be a constant. For unsteady flow, (28) is no longer valid, but (29) is. For example, at the instant when $Q = 0$, v_r is also zero, but Γ is obviously not because of *vortex inertia* (i.e. even without a flow rate, the liquid continues to rotate). For an inviscid liquid, such a flow will continue forever with maximum Γ. For real flow, if the viscous time ($\sim R^2/v$) is much larger than the period of oscillation, we can assume that (30) is correct for the maximum Q, and thus $\Gamma = Q_{max}/\theta$.

Using (30) and (12) we find the important relationship between system dynamics and design parameter θ

$$|\dot{y}|_{max} = (nb/2\pi R) = \theta. \tag{31}$$

This equation is the condition for the swirler to be at rest. The derivation of (31) is speculative and needs to be verified by experiment with unsteady separation process. Equation (29) assumes an axisymmetric configuration of the entrances. So the number of entrances (n) should be relatively large.

Semi-Otto cycle: Experimental realization of the Otto cycle is difficult because we need a very high maximum pressure in order to use a realistic compression ratio; our goal is only to prove the concept, i.e. to establish the possibility of oscillating stable cavity generation. In order to model the cavity with comparable liquid velocities, we are going to use a *semi-Otto cycle* (as shown in figure 8) in our subscale "cold" experimental model (using compressed air supply) because it operates at lower pressures than the Otto cycle at the same compression ratio.

In our experiments the parameter E is not meaningful, therefore we introduce a pressure ratio parameter χ which allows us to consider both cycles:

$$\chi = p_{max}/p_{min} = p_c/p_a. \tag{32}$$

The values E and χ are related. For the Otto cycle,

$$\chi = \epsilon^k E. \tag{33}$$

Introducing χ into (27), we obtain

$$f_1 - f_2 = \left(\frac{\sigma + y_0}{\sigma + y}\right)^k - \chi \left(\frac{\sigma - y_0}{\sigma + y}\right)^k. \tag{34}$$

For the semi-Otto cycle we have

$$f_1 - f_2 = 1 - \chi \left(\frac{\sigma - y_0}{\sigma - y} \right)^k. \tag{35}$$

3.4 Energy equation and operating characteristics

We found that (13) needs to be solved for half a period $0 \leq \tau \leq \tau_c/2$ only, where $\dot{y}(\tau) \leq 0$ and $(\zeta/2)|\dot{y}|\dot{y} = -(\zeta/2)\dot{y}^2$. Then it can be shown that (13) can be written in the form

$$\frac{1}{2} \frac{d}{dy} \left[(\dot{y}^2 + 1)\ln \frac{1}{\sigma^2 - y^2} \right] = \frac{\zeta}{2} \dot{y}^2 - \Pi(f_2 - f_1). \tag{36}$$

Integrating (36) for the half-period of oscillation between $y = -y_0$ and $y = y_0$, we obtain the energy equation

$$\frac{\zeta}{2} \int_{-y_0}^{y_0} \dot{y}^2 dy = \Pi \int_{-y_0}^{y_0} (f_2 - f_1) dy. \tag{37}$$

Equation (37) states that inertial features do not participate directly in the energy balance. The right-hand side of (37) is the work (during a thermodynamic cycle) and the left-hand side is the energy consumption. Equation (37) helps to deduce the power W of the engine. The general expression for W is

$$W = 2\pi R^2 h \rho V^2 f \Pi \int_{-y_0}^{y_0} (f_1 - f_2) dy, \tag{38}$$

where f is the frequency,

$$f = 2V/R\tau_c. \tag{39}$$

For an Otto cycle, a simple calculation gives

$$\int_{-y_0}^{y_0} (f_1 - f_2) dy = \frac{\sigma + y_0}{k - 1} \eta E_0 \tag{40}$$

where $\eta = 1 - \epsilon^{1-k}$ is the well-known efficiency for Otto cycle (Faires & Virgil 1970), ϵ and E_0 are given by (19) and (26). For the semi-Otto cycle, using (27), we have

$$\int_{-y_0}^{y_0} (f_1 - f_2) dy = \chi \frac{\sigma - y_0}{k - 1} (1 - \epsilon^{1-k}) - 2y_0. \tag{41}$$

It can be shown that the efficiency for this semi-Otto cycle is

$$\eta = 1 - \frac{k(\epsilon - 1) + \chi \epsilon^{1-k} - \epsilon}{\chi - 1}. \tag{42}$$

The efficiency η of the Otto cycle depends on the compression ratio ϵ alone while that of the semi-Otto cycle depends on both ϵ and χ. Note that at a fixed χ, the efficiency (42) has a maximum at $\epsilon = \chi^{1/k}$, where $\partial \eta / \partial \epsilon = 0$. Then

$$\eta_{max} = 1 - \frac{k(\chi^{(1/k)} - 1)}{\chi - 1} \quad \text{or} \quad \eta_{max} = 1 - \frac{k(\epsilon - 1)}{\epsilon^k - 1}. \tag{43}$$

Figure 9 shows a comparison of this efficiency with that of the Otto cycle with $k = 1.4$.

Note that for any given compression ratio ϵ, $\eta_{Otto} > \eta_{max}$. However, for a fixed pressure ratio χ in the working range, $\eta_{max} > \eta_{Otto}$.
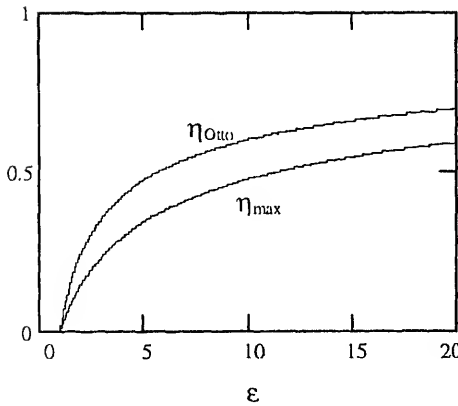
Figure 9. Comparison of efficiency.

3.5 Free oscillations

Unlike the conventional solid piston engines where the piston displacements are fixed, the VLPE has a variable displacement which depends on system parameters. In order to achieve the maximum oscillating amplitude (which gives maximum power and efficiency), it is very important to know the eigenfrequency of the VLPE. This may also benefit the creation of a "shrinking engine" (*Popular mechanics* 1995) whose cylinder size can be changed corresponding to load to reduce energy losses.

Small oscillations: The eigenfrequency for small free oscillations can be obtained from (13) by assuming that both y and $\dot{y}$ are small; that is

$$\ddot{y} \ln \frac{1}{\sigma^2} + \frac{1}{\sigma^2} y = \Pi(f_1 - f_2). \tag{44}$$

For an adiabatic process, by letting $S = 0$ in (18), we have for the right-hand side of (44),

$$\Pi(f_1 - f_2) = \Pi\left[\left(\frac{V_0}{V_1}\right)^k - \left(\frac{V_0}{V_2}\right)^k\right] = \Pi\left[\left(\frac{\sigma + y_0}{\sigma + y}\right)^k - \left(\frac{\sigma + y_0}{\sigma - y}\right)^k\right]. \tag{45}$$

When $y \ll 1$ and $y_0 \ll 1$,

$$\left(\frac{\sigma + y_0}{\sigma + y}\right)^k = 1 - \frac{k}{\sigma} y + \cdots \quad \text{and} \quad \left(\frac{\sigma + y_0}{\sigma - y}\right)^k = 1 + \frac{k}{\sigma} y + \cdots \tag{46}$$

Using these expansions in (44) and (45), the equation for small free oscillations becomes

$$\ddot{y} \ln \frac{1}{\sigma^2} + \left(\frac{1}{\sigma^2} + \frac{2k\Pi}{\sigma}\right) y = 0, \tag{47}$$

and the eigenfrequency is

$$\omega_0 = \left(\frac{1 + 2k\Pi\sigma}{\sigma^2 \ln(1/\sigma^2)}\right)^{1/2}. \tag{48}$$
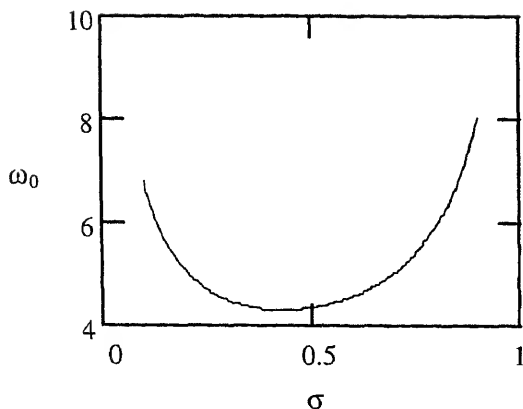
Figure 10. ω_0 as a function of σ with fixed Π.

(48) shows that ω_0 has a minimum value with respect to σ. This value is achieved when σ satisfies the following equation

$$(1 + \ln \sigma)(1 + k\Pi\sigma) = \tfrac{1}{2}.$$
(49)

Figure 10 depicts ω_0 as a function of σ with $k = 1.4$ (for air) and nondimensional pressure $\Pi = 4.0$ (for $p_0 = 10^5$ Pa and $V = 5$ m/s, $\Pi = p_0/\rho V^2$). From this figure, we can see that ω_0 has a minimum at $\sigma = \sigma_{min}$. This unique feature will be explained later together with figure 14 below. Figure 11 shows that σ_{min} has its largest value, $\sigma_{min} = (1/e)^{1/2} \approx 0.6065$, at $\Pi = 0$ and decreases with increasing Π. Figure 12 shows how the minimum frequency changes with Π.

Nonlinear free oscillations: Let us consider (36) with $\zeta = \Pi = 0$. In this case, (36) can be integrated to give

$$(\dot{y}^2 + 1) \ln \frac{1}{\sigma^2 - y^2} = \text{const.}$$

The constant can be found from the condition $\dot{y}(y_0) = 0$, and solving for $\dot{y}$ gives

$$\dot{y} = \sqrt{\frac{\ln(\sigma^2 - y_0^2)}{\ln(\sigma^2 - y^2)} - 1}.$$
(50)
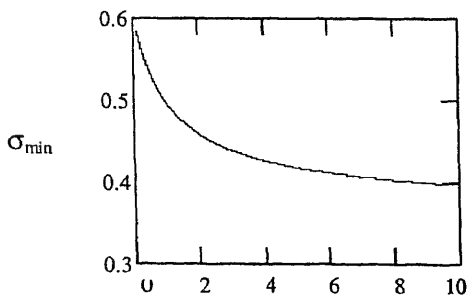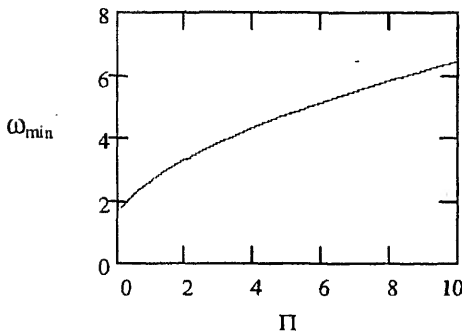
Figure 12. ω_{min} as a function of Π.

Integrating (50) for half-period $\tau_c/2$ we have

$$\tau_c = 4 \int_0^{y_0} \left(\frac{\ln(\sigma^2 - y_0^2)}{\ln(\sigma^2 - y^2)} - 1 \right)^{-1/2} dy. \qquad (51)$$

Defining the eigenfrequency f_c as

$$f_c = 1/\tau_c, \qquad (52)$$

from (39) we have

$$f = (2V/R) f_c. \qquad (53)$$

Figure 13 shows f_c as a function of y_0 for the cases $\sigma = 0.4, 0.5$, and 0.6. We see that for the typical σ range, eigenfrequency increases with increasing amount of fluid in the system as well as with oscillating amplitude y_0. Both these features are the opposite of those of an ordinary oscillating system like a U-tube where increasing amount of liquid decreases the eigenfrequency and oscillating amplitude decreases with increase in eigenfrequency. These features are useful for a real engine, because with the increase of engine frequency and power, the oscillating amplitude increases and so does the efficiency, (19).

Figure 14 shows f_c as a function of σ for the case $y_0 = 0.15$. The eigenfrequency has a minimum with respect to the gas/liquid ratio σ for finite amplitude oscillation. This somewhat unusual non-monotonic behaviour of the eigenfrequency is related to the strong dependence of centrifugal force $\rho v_\varphi^2 / r$ on r and can be explained by making analogy with
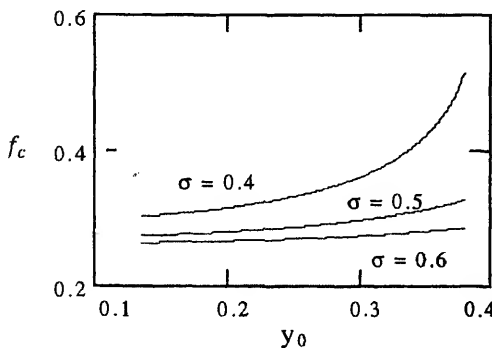
Figure 13. f_c as a function of y_0 with fixed σ.
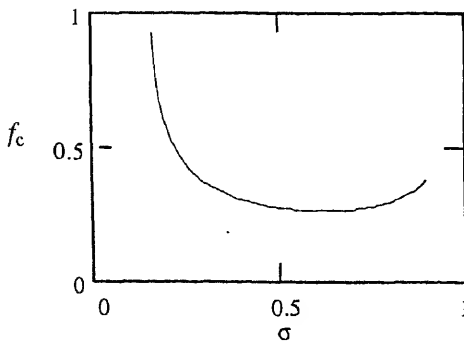
Figure 14. f_c as a function of σ with fixed y_0.

a U-tube system. For the U-tube, the eigenfrequency is $\omega = (g/l)^{1/2}$, where l is the height. For VLPE, when σ is small, we have larger amount of liquid such that r is small. Then the centrifugal force $\rho v_\varphi^2 / r$ tends to infinity as σ tends to zero. This is equivalent to a larger g case for the U-tube. Therefore, the eigenfrequency is very high for small σ. On the other hand, when σ is large, we have small amount of liquid such that the corresponding l is small. Between these two extremes there must be a minimum.

3.6 Stability criterion

The stability problem is very complicated in the general case, because the corresponding linearized hydrodynamic equation has time-dependent coefficients, and Floquet theory must be used. Therefore, we prefer to investigate the general stability problem by experiment. However, a simple criterion S_t, necessary for stability, can be found from elementary considerations:

$$S_t = \min(g_c/g_r) = \min 1/[2(\sigma + y)\ddot{y} - \dot{y}^2], \tag{54}$$

where $g_c (= v_\varphi^2/r)$ is the centrifugal acceleration and g_r is the radial acceleration. If the maximum is achieved at the moment of maximum compression, (54) is slightly simplified ($\dot{y} = 0$). Clearly, g_c must be greater than g_r or else the cavity will collapse; hence, the minimal $S_t = 1$. We therefore use $S_t > 1$ as the stability criterion. The larger the value of S_t, the more stable the rotating flow. Another related condition is $g_c \gg g$, where g is gravity.

3.7 Variable circulation

The previous formulation was based on the assumption that $\Gamma = $ const. But how do we choose this constant? It clearly depends on flow rate Q; in particular, if $Q \equiv 0$ then $\Gamma \equiv 0$. For a steady flow, $\Gamma = -\theta Q$, where $\theta = v_\phi/v_r = \tan \alpha = $ const is a swirl parameter and α is the blade angle of the swirler device. This follows from the Kutta condition (i.e. no separation) for inviscid flow past blades. If a process is unsteady, the direct application of the Kutta condition should give

$$\Gamma(t) = -\theta Q(t), \quad Q(t) < 0. \tag{55}$$

But the validity of the Kutta condition for unsteady flow is doubtful because of possible flow separation. Nevertheless, in this section, we consider (55) to be valid. For such a situation, Γ is not a constant and depends not only on t, but also on r. We use an equation for Γ

$$\frac{\partial \Gamma}{\partial t} + v_r \frac{\partial \Gamma}{\partial r} = 0; \quad v_r = \frac{Q}{r}. \tag{56}$$

Although $\Gamma = $ const. is a solution of (56), it does not satisfy (55). (56) can be solved by the method of characteristics. Let us introduce the variables $z_i(\tau)$ and $s(t)$, such that

$$z_i(\tau) = \frac{r_i^2(\tau)}{R^2}, \quad i = 1, 2; \quad ds = \frac{2Q}{R^2} dt. \tag{57}$$

Note that if $Q = $ const. $= RV$, we have $s = \tau$.

From (10),

$$r_1^2 + r_2^2 = 2R^2 \sigma. \tag{58}$$

Using the new variables, (56) can be rewritten as

$$\frac{1}{R^2} \frac{\partial \Gamma}{\partial s} + \frac{\partial \Gamma}{\partial r^2} = 0.$$

The general solution of this equation is

$$\Gamma(r, \tau) = F(r^2 - R^2 s), \quad \cdot \tag{59}$$

where F is an arbitrary function which can be found from the boundary condition at $r = R$. The boundary condition (55) then becomes

$$F(R^2(1 - s)) = -\theta Q(\tau).$$

If s is known, we can determine the function F using the inverse function $\tau = \tau(s)$.

To derive the equations of motion, we assume that at the initial time $\tau = 0$ both cylinders are filled with equal amounts of liquid, rotating with $\Gamma = \Gamma_0 = $ const. Once the motion starts, a part of the liquid flows into one cylinder from the other according to (55). Under such conditions, liquid in the first cylinder still has circulation Γ_0. Let b_1 be the radius of the boundary between this liquid and the remainder liquid having another circulation.

Thus the circulation distribution is

$$\Gamma(r, t) = \begin{cases} \Gamma_0, & r_1 \leq r \leq b_1, \\ F(r^2 - R^2 s), & b_1 < r \leq R. \end{cases}$$

Mass balance gives

$$\pi(R^2 - b_1^2) = \pi(R^2 - r_1^2) - \pi(R^2 - r_{1\max}^2).$$

$$\int_{r_1}^{R} \frac{\Gamma_0^2}{r^3}\, dr = 2\pi^2 c^2 \gamma_0^2 \left(\frac{1}{z} - \frac{1}{1 + z - z_{max}}\right). \tag{61}$$

Consider the integral

$$\int_{r_1}^{b_1} \frac{\Gamma_0^2}{r^3}\, dr + \int_{b_1}^{R} \frac{F^2(r^2 - R^2 z)}{r^3}\, dr. \tag{62}$$

The first part of (62) is given by (61). To calculate the second part, we introduce variables

$$z' = 1 - r^2/R^2; \quad z_* = 1 - b_1^2/R^2; \quad \xi = z + z'.$$

Using (59), the second part of (62) becomes

$$\int_{b_1}^{R} \frac{F^2(r^2 - R^2 z)}{r^3}\, dr = \int_0^{z_*} \frac{F^2(R^2(1 - z - z'))}{2R^2(1 - z')^2}\, dz'$$

$$= \int_z^{z_{max}} \frac{F^2(R^2(1 - \xi))}{2R^2(1 + z - \xi)^2} = \frac{\theta^2}{2R^2} \int_z^{z_{max}} \frac{Q^2(\tau(\xi))}{(1 + z - \xi)^2}\, d\xi \tag{63}$$

$$= \frac{\theta^2 C^2}{8} \int_{z(\tau)}^{z_{max}(\tau)} \frac{\dot{z}^2(\sigma)}{[1 + z(\tau) - z']^2}\, dz'.$$

In order to integrate equation (63), it is sufficient to know the dependence of $\dot{z}(z)$ for all previous time. In the same way, we can obtain the corresponding relation for the second cylinder by integrating (2) from r_2 to R and taking into account the boundary conditions determined by the swirler at the moment of inflow. Thus, we obtain

$$\Gamma_2(R, t) = F_2(R^2(1 - s)) = \theta Q(\tau(s)), \quad Q > 0 (Q_2 = -Q).$$

The result is slightly different from (8).

Finally we obtain the generalization of (13):

$$\ddot{z} \ln \frac{1}{z(2\sigma - z)} + \left(1 + \frac{\dot{z}^2}{4}\right)\left(\frac{1}{2\sigma - z} - \frac{1}{z}\right) + \frac{1}{1 + z - z_{max}}$$

$$- \frac{1}{1 + z_{min} - z} + \frac{\theta^2}{4}\left[\int_{z_{min}}^{z} \frac{\dot{z}^2(z')}{(1 - z + z')^2}\, dz'\right. \tag{64}$$

$$\left. - \int_z^{z_{max}} \frac{\dot{z}^2(z')}{(1 - z + z')^2}\, dz' + \zeta \frac{\dot{z}^2}{4} \operatorname{sign} \dot{z} - 4\sigma \frac{\sigma - z}{z(2\sigma - z)}\right] = 0,$$

where $z_{max}(t) = \max_{0 \le \tau' \le \tau} z(t')$, $z_{min}(t) = \min_{0 \le \tau' \le \tau} z(t')$, and $0 \le z(t) \le \min(1, 2\sigma)$.

The integro-differential equation (64) determines the evolution $s(\tau)$ for a given variation of the temperature in the cavities.

The initial conditions are

$$z(0) = z_0, \quad \dot{z}(0) = 0, \quad z_{max}(0) = z_0, \quad \text{and} \quad z_{min}(0) = z_0. \tag{65}$$
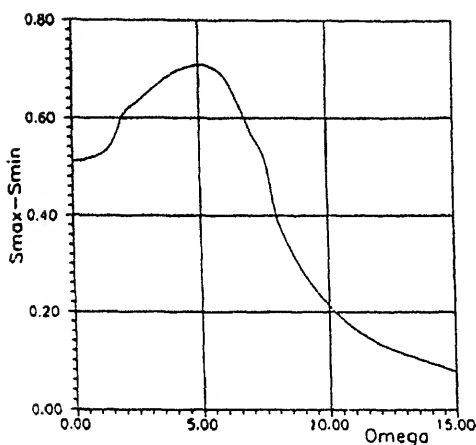
Figure 15. Resonance frequency.

Numerical calculations of the system (64)–(65) were performed using the Runge–Kutta method with adaptive step control. The integrals in (64) were calculated using the technique of nonlocal splines at each step of integration.

Figure 15 shows the dependence of amplitude $\Delta S = S_{max} - S_{min}$ for the periodic regime versus the forcing frequency ω. This shows that there is some resonance frequency ω^*. The small bumps correspond to $\omega^*/2$ and $3\omega^*/2$. The solution near ω^* has a minimal transition time to periodic motion. The phase portrait of the transient process for $\omega = \omega^*$ is represented in figure 16. The phase trajectories of the transient processes are shown in figure 17 for $\omega > \omega^*$ and in figure 18 for $\omega < \omega^*$.

Note that these results suggest that we can expect the minimal perturbations of cavity surfaces near resonance. The complex transient process with high-frequency oscillations corresponds to a flow with $\omega \ll \omega^*$ (figures 19 and 20).

If $\omega \gg \omega^*$, the high-frequency perturbations do not appear but amplitudes z and $\dot{z}$ grow strongly and exceed amplitudes of periodic regime (figures 21 and 22). This analysis shows that regime with $\omega \approx \omega^*$ is preferable.
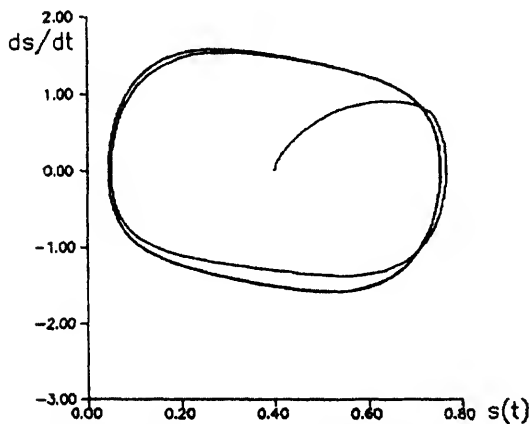
Figure 16. Phase portrait of the transient process for $\omega = \omega$.
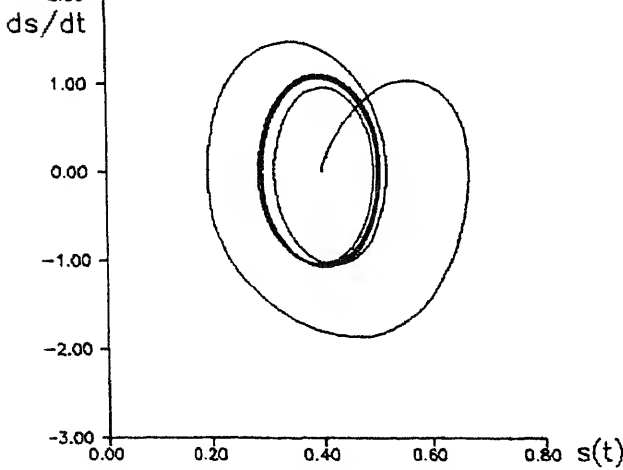
The calculations with $\Gamma = $ const. show that an important difference exists only for small frequencies $\omega \ll \omega^*$. If $\omega > \omega^*$, the resonance curves practically coincide (the resonance curve for $\Gamma = $ const. is shown in figure 23 by the dashed curve).

The general conclusion is: if the frequency is sufficiently high ($\omega > \omega^*$), the assumption of $\Gamma = $ const. is reasonable, because, in this case, the transient processes can be considered quasi-stationary processes.

3.8 The role of viscosity

Let the flow rate $Q(t)$ be given. The equation for circulation Γ is

$$\frac{\partial \Gamma}{\partial t} + \frac{Q(t)}{r}\frac{\partial \Gamma}{\partial r} = \nu r \frac{\partial}{\partial r}\frac{1}{r}\frac{\partial \Gamma}{\partial r},\tag{66}$$
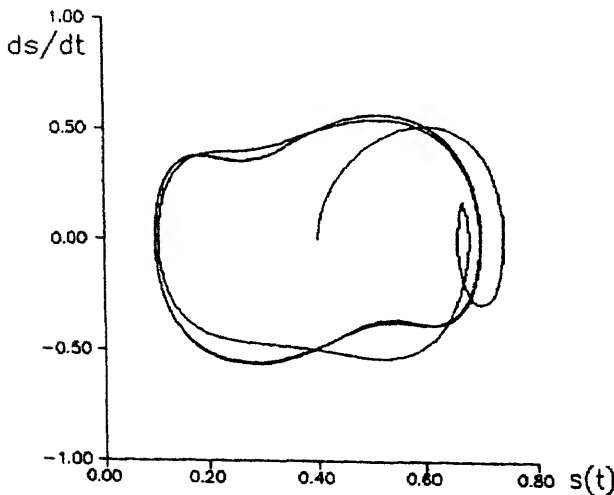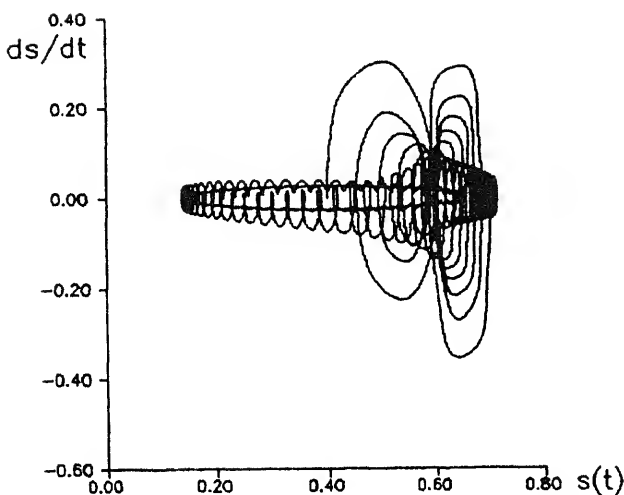
Figure 19. Complex transient process for $\omega \ll \omega^*$.

where ν is the kinematic viscosity. Assume $r_1 = r_0(1 + a \sin \omega t)$, and choose R as the length scale to nondimensionalize r. Introducing other variables

$$\tau = \frac{\nu t}{R^2}; \quad \Omega = \frac{\omega R^2}{\nu}; \quad q = \frac{Q(t)}{\nu}; \quad \text{and} \quad q_0 = \Omega r_0^2 a,$$

we have

$$Q = R^2 r_1 \dot{r}_1 = R^2 r_0^2 a \omega (1 + a \sin \Omega \tau) \cos \Omega \dot{\tau},$$

and

$$q = q_0(1 + a \sin \Omega \tau) \cos \Omega \tau.$$
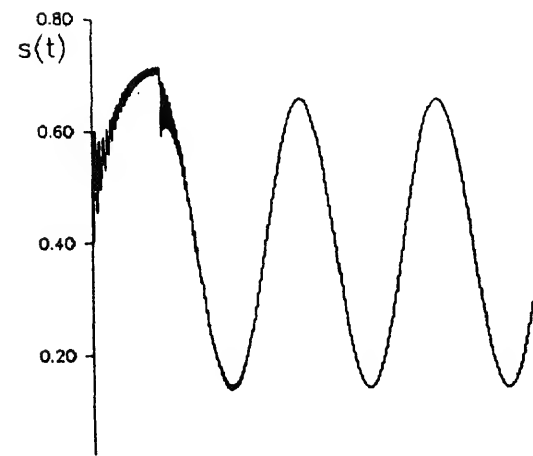
Figure 20. Complex transient process for
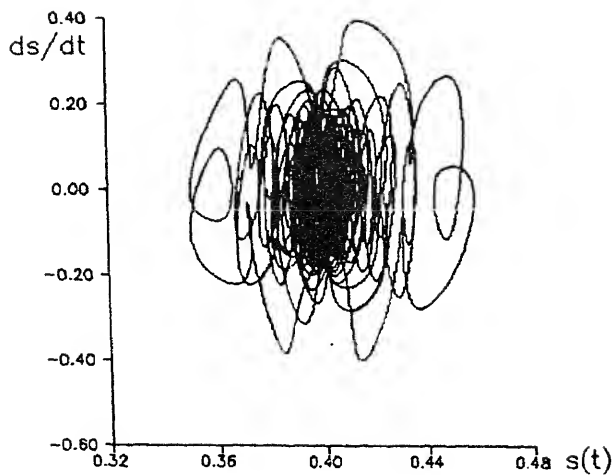
Figure 21. Transient process for $\omega \gg \omega^*$.
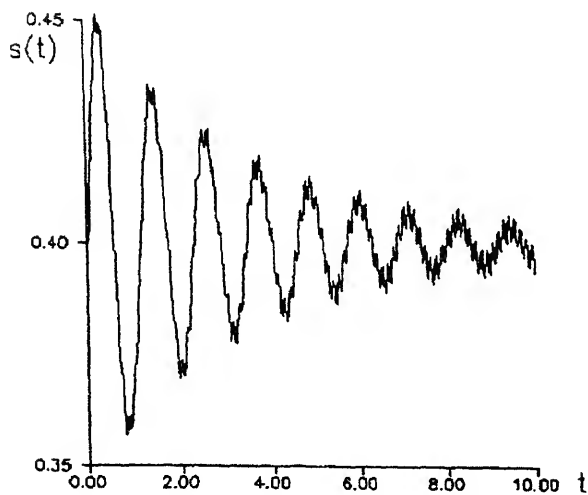
Figure 22. Transient process for $\omega \gg \omega^*$.
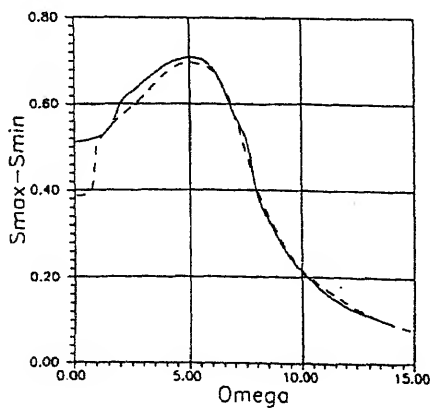
Figure 23. Comparison of resonance curves.

Equation (66) is transformed into

$$\frac{\partial \Gamma}{\partial \tau} = \frac{\partial^2 \Gamma}{\partial r^2} - \frac{1 + q(\tau)}{r} \frac{\partial \Gamma}{\partial r}, \qquad r_1 \leq r \leq 1$$

with the boundary conditions:

$$\Gamma(\tau) = 1; \qquad \frac{\partial \Gamma}{\partial r} = \frac{2\Gamma}{r} \qquad \text{at} \quad r = r_1.$$

The last one is the free surface condition.

Let us consider two cases: $\Omega \ll 1$ and $\Omega \gg 1$. For the first case

$$\Gamma = [r^{q+2} + (q/2)r_1^{q+2}]/[1 + (q/2)r_1^{q+2}].$$

If $q = 0$, $\Gamma = r^2$ corresponding to solid body rotation. If $q \gg 1$, $\Gamma = r^{q+2}$, almost no rotation. If $q \ll -1$, $\Gamma = 1 - 2/|q|(r/r_1)^{q+2}$ and $\Gamma \approx 1$. For $\Omega \gg 1$, $\Gamma \equiv 1$. These results justify formulation of the problem using $\Gamma = \text{const.}$, which is correct for inviscid flow, for sufficiently high frequency, which is the most interesting case for the real LPE.

4. Numerical analysis of VLPE

We first want to estimate the engine parameters using an Otto-cycle. First, we estimate the natural oscillation frequency. Using (48), for typical values of the parameters, viz., $k = 1.4$, $\sigma = 0.25$, and $\Pi = 4$, we have $\omega_0 = 6.26$ and $\omega = 626 \, \text{rad/s}$. This value of ω corresponds to $n \approx 100 \, \text{rev/s} \approx 6000 \, \text{rpm}$, which is typical for the maximum rpm of an internal combustion engine for automobiles. Thus, the vortex engine has good dynamic characteristics and can be started quickly.

The value of ζ is roughly estimated as follows. For the mean values we have $W = \Delta p \, Q_v$, or from (2) and (9)

$$W = \zeta(\rho Q_v^3)/(8\pi^2 R^2 h^2). \tag{67}$$

Let $W = 100 \, \text{kW}$ ($= 134 \, \text{hp}$), $R = 0.1 \, \text{m}$, $h = 0.1 \, \text{m}$, and $p = 1000 \, \text{kg/m}^3$. To estimate the value Q_v, we can use the formula $Q_v = \pi r^2 h n$, where r is the mean radius which can be found by setting $r_1 = r_2$. Using (3) we find $r = R\sqrt{\sigma}$ and for $\sigma = 0.25$, we have $r = 0.5R = 0.05 \, \text{m}$. If $n = 100 \, \text{rps}$, $Q_v = 0.078 \, \text{m}^3/\text{s}$. From (67) then

$$\zeta = (8\pi^2 R^2 h^2 W)/(\rho Q_v^3) = 1664.$$

For the given dimensions and flow rate Q_v, we estimate the fluid velocity $V = 5 \, \text{m/s}$ at the entrance of the cylinder where $r = R$. This is a moderate velocity.

For numerical solution of (13), we use condition (20) and integrate for half a period until $\dot{y} = 0$. At this moment, if (22) is satisfied, we have y_0 and all other engine characteristics. Otherwise, we use secant or Newton's method to find a new y_0 and start again until (22) is satisfied.

For the example considered (i.e. $W = 100 \, \text{kW} = 134 \, \text{hp}$, $R = h = 0.1 \, \text{m}$, $\zeta = 5000$, $V = 5 \, \text{m/s}$, and $\sigma = 0.4$), we obtained numerically $\epsilon \approx 10$, $\eta = 60\%$, and a large margin for stability, $S_t = 72$, corresponding to the acceleration $a_c \approx 1300 \, \text{g}$ at the maximum compression point. In order to have an acceptable efficiency value, we assign
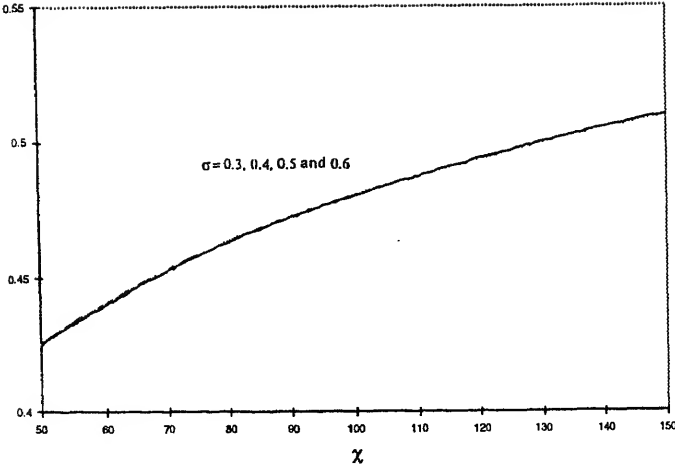
Figure 24. Efficiency for Otto cycle with fixed ζ.

the compression ratio, (19). We want to have a maximum magnitude of oscillation, in which case the maximum radius $r_1 = R$. This yields $y_0 = 1 - \sigma$, which together with (19) gives $\sigma = 0.55$ and $y_0 = 0.45$.

We chose the following parameters for the experimental device: $R = 0.1\,\text{m}$, $b = 0.019\,\text{m}$, $h = 0.3\,\text{m}$ and $v_\varphi = 5\,\text{m/s}$. From numerical calculations, $|\dot{y}|_{\max} = 0.03$. The corresponding working parameters are: $\beta = 4.49$, $\zeta = 20\,000$, $f = 1.05\,\text{Hz}$, $a_c = 806\,\text{g}$, $S_t = 1986$, $W = 669\,\text{kW}$, $\Delta p = 1.12\,\text{atm}$, $Q_{\max} = 0.028\,\text{m}^3/\text{s}$, $p_{R1} - p_{R2} = 2.25\,\text{atm}$, $p_{R1} = 2.25\,\text{atm}$, and $p_{R1} - p_{r1} = 1.13\,\text{atm}$. The low frequency here is only for detailed investigation of the process and is not typical for the vortex engine. All these parameters are used and verified in our experiments. A distinct feature of the vortex liquid piston engine
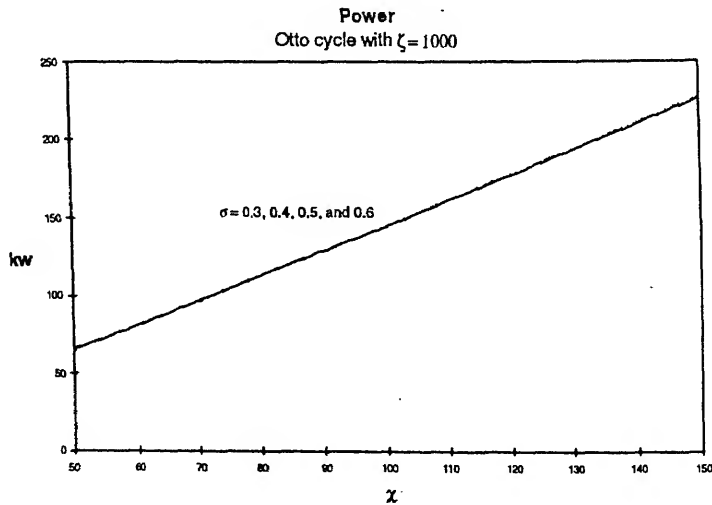
Figure 25. Power for Otto cycle with fixed ζ.
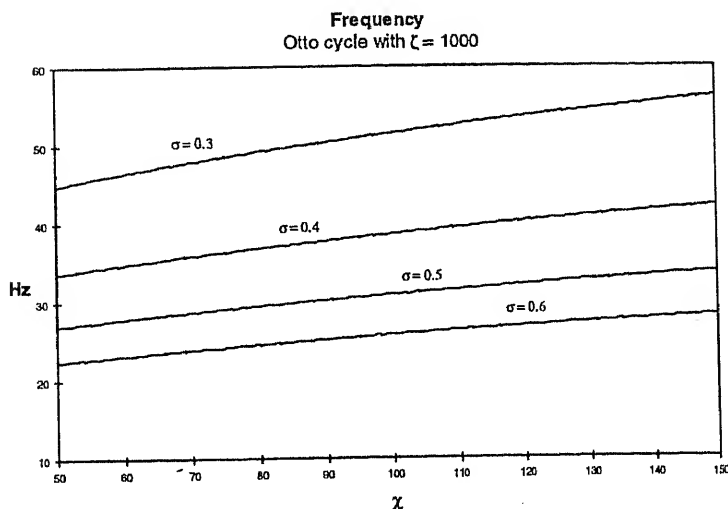
Figure 26.   Frequency for Otto cycle with fixed ζ.

is that its efficiency increases with increasing power. This can be explained by the increase of the oscillation amplitude of the radii of the cavities and therefore of the compression ratio ϵ. The vortex engine differs significantly from the conventional internal combustion engine where ϵ is fixed and η decreases with increasing power.

To complete the numerical analysis, we calculated the power, efficiency, and frequency of this engine for Otto cycle. Figures 24–26 show the results for a constant $\zeta = 1000$. From these figures, we can see for the Otto cycle:

(1) both engine power and efficiency increase with χ, but change little with σ.

(2) engine frequency decreases with σ (figure 26). This is a unique feature because it shows that the frequency increases with the mass of the oscillating system.
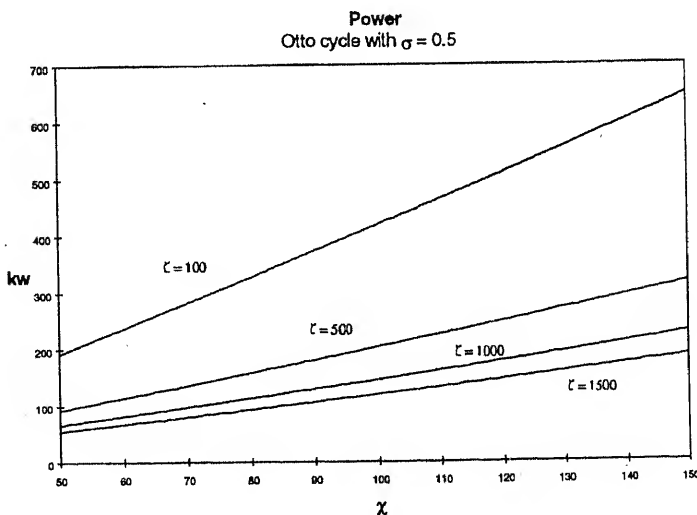
Figure 27.   Power for Otto cycle with fixed σ.
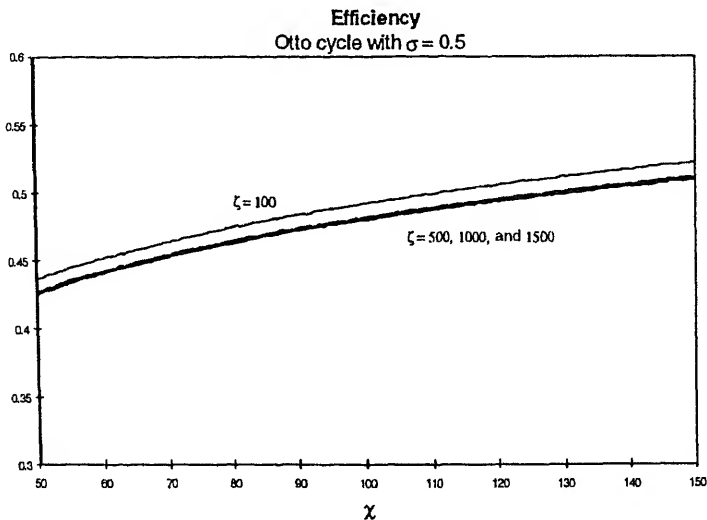
Figure 28.   Efficiency for Otto cycle with fixed σ.

Figures 27–29 show the results with a constant $\sigma = 0.5$. From these figures, we can see for the Otto cycle that power, efficiency, and frequency, all decrease with increasing ζ.

Figures 30 and 31 show the case when χ is small. From these we can see for the semi-Otto cycle:

(1) engine power increases with increasing σ (figure 30b) and χ (figure 30a);

(2) engine power has a maximum with respect to $\zeta(\zeta_{\max w})$ (figures 30a and b); with increase of σ (less fluid), $\zeta_{\max w}$ decreases (figure 30b); however, if χ is large enough, $\zeta_{\max w}$ is almost constant with increasing χ;
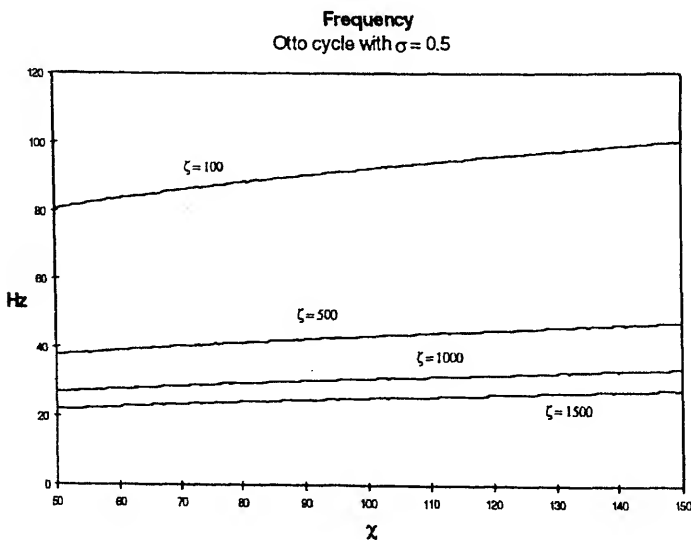
Figure 29.   Frequency for Otto cycle with fixed σ.

Power as a function of ζ and χ
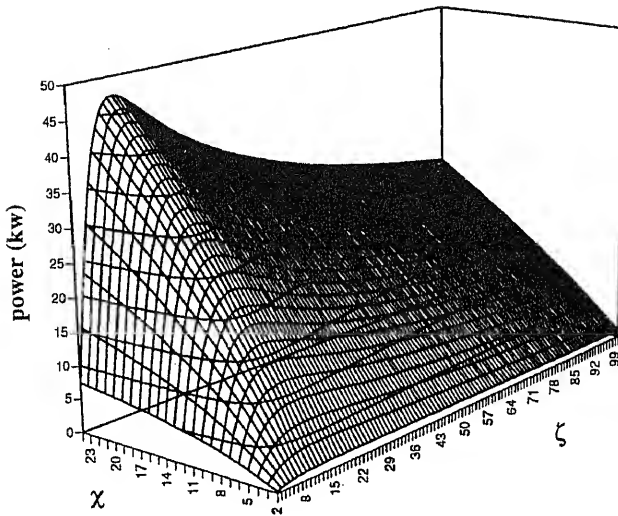
a)

Power as a function of resistance
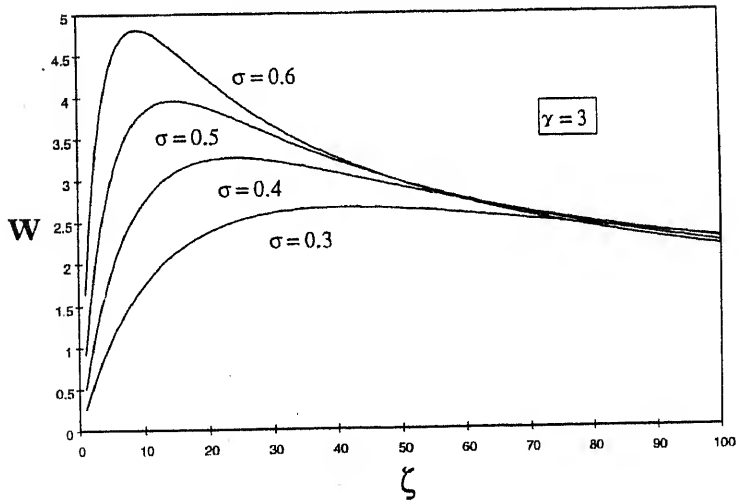
b)

Figure 30.   Power for the semi-Otto cycle.

Efficency as a function of ζ and χ

a)
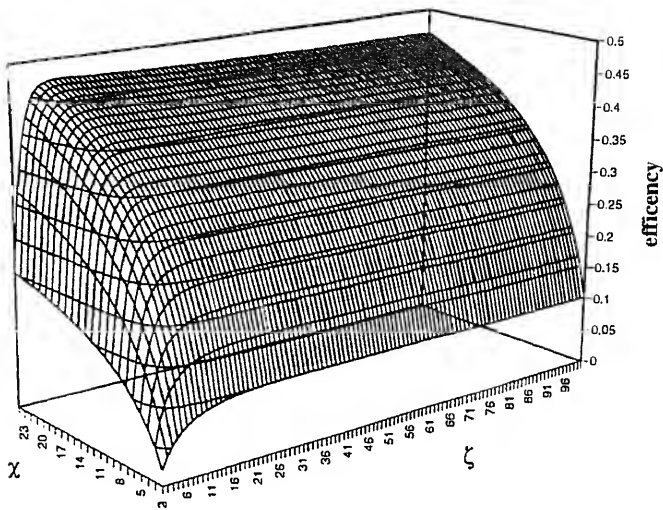
Efficency as a function of resistance
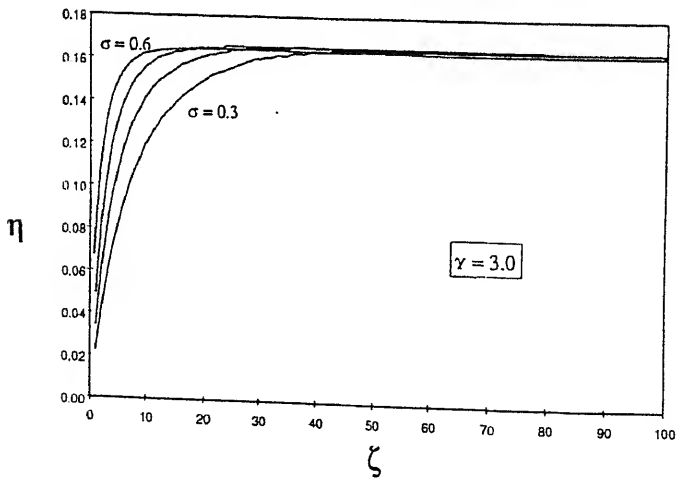
b)

Figure 31. Efficiency for the semi-Otto cycle.

Power as a function of ζ and χ
Otto cycle

$\sigma = 0.5$

a)

Efficency as a function of ζ and χ
Otto cycle

$\sigma = 0.5$
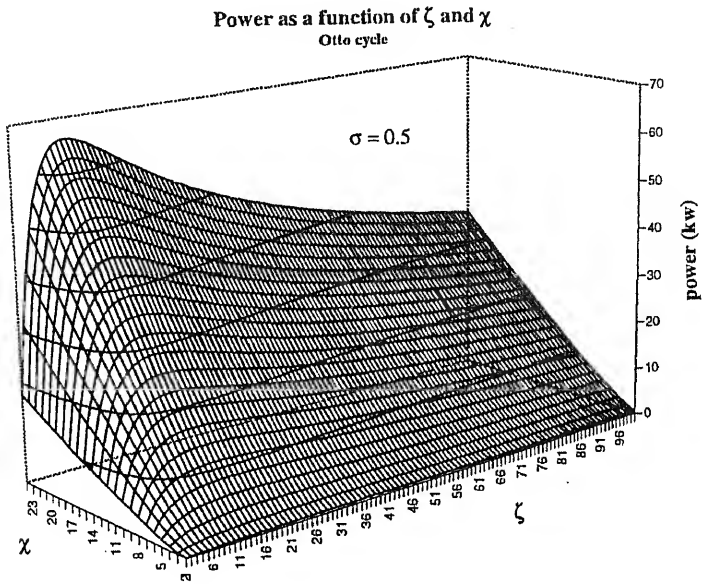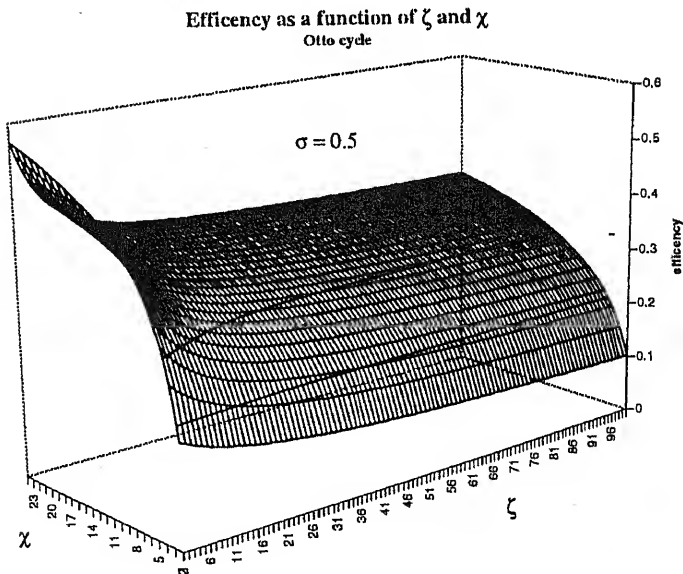
b)

Figure 32. Power and efficiency for Otto cycle.

(3) engine efficiency increases with increasing σ (figure 31b) and χ (figure 31a);

(4) engine efficiency has a maximum with respect to ζ ($\zeta_{\max_\eta}$) (figure 31a); with increase of σ (less fluid), $\zeta_{\max_\eta}$ increases (figure 31b); however, when ζ is large enough, engine efficiency will be almost constant (figure 31a).

Figure 32 shows the similar case for the Otto cycle.

5.   Engine design and experimental analysis

After finding all the necessary relationships and the optimized design parameters, we built a working model using water and air. Figure 33 shows our first design. It consists of two cylinders connected by two tangential channels and a check valve, so that water can flow in each channel only in one direction.

We used an external oscillating circuit to control all four solenoid valves (one air inlet and one exhaust valve for each of the chambers). The control circuit is shown in figure 34, where $R2$ can be adjusted to achieve different oscillation frequencies.

After the first prototype was built and tested, we discovered a serious problem with this design. We had assumed that high-speed water would enter the chamber and rotate inside it to create a stable axisymmetric cavity. However, a very complicated turbulent flow without swirl and cavity was produced (figure 35) because of the tangential outlet whose lip generates a big disturbance due to flow impingement and thus an obstacle to the rotating flow. A swirling wall jet of water hits the outlet conduit and is deflected from this obstacle back to the centre of the chamber.

In order to eliminate this "outlet effect" and provide a smooth surface for the rotating flow, we completely blocked the upper channel (figure 36a) and divided the lower one into two halves (figure 36b): one half is for incoming flow and the other half for outflow. This time swirling flow and a cavity were created. However, on injection of pressurized air into the cavity, the cavity immediately collapsed. After careful analysis, we found the main
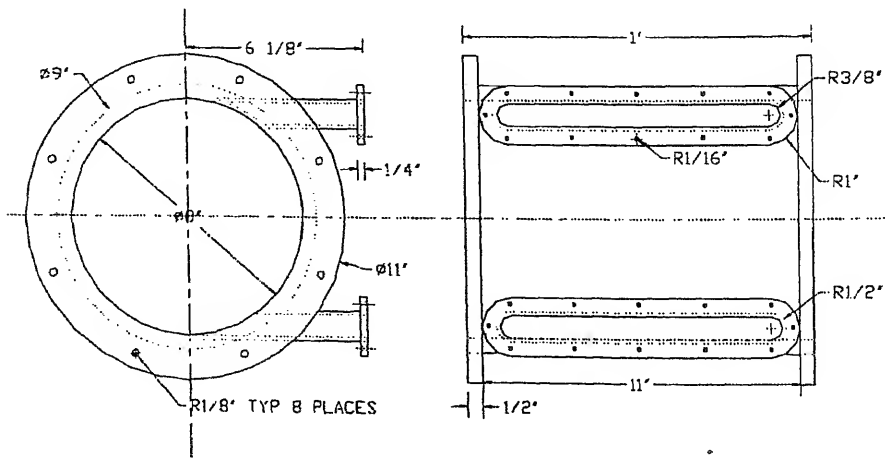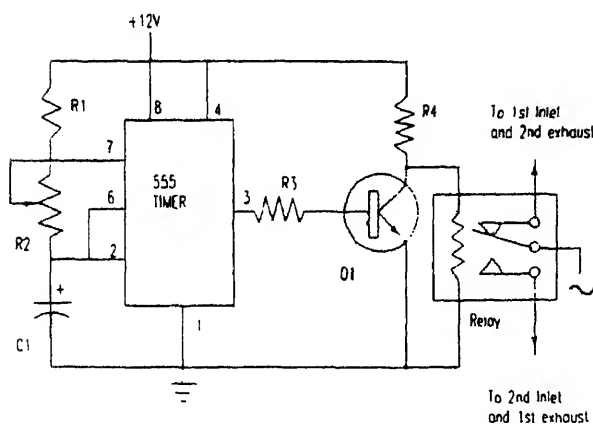
Figure 34.   Control circuit.

reason is because the outlet is not axisymmetric. This breaking of the system symmetry causes the cavity to be destroyed.

Since axisymmetry is crucial, it was thus necessary to modify our model. Figure 37 shows a meridional cut of the actual modification where the block in the middle prevents air from exiting with the liquid when this model works horizontally.

Figure 38 shows a picture of the experimental configuration. As expected, the cavity created with this modified model expands smoothly when pressure is applied. Therefore, we have achieved a working chamber with an expandable and contractible cavity inside it. The cavity created is quite stable as predicted by our theory. All the system parameters measured, such as oscillation frequency, flow rate, and power, are comparable to our theoretical model and numerical results. Figure 39 shows the minimum cavity in the periodic working engine. One can see from these pictures that the cavity, though successfully generated, is not quite axisymmetric. The reason is the non-axisymmetry of the inlets and outlets of the system and machining roughness. This is why further redesign was necessary.

We realized that the most important design goal is to make the system as axisymmetric as possible, because any non-axisymmetry will most likely cause the cavity to be unstable or to collapse under sudden application of pressure. In order to have an axisymmetric inlet, we decided to use a distributor as shown below (figure 40). This distribution forms evenly distributed water jets and helps us create a more stable cavity.

Flow in
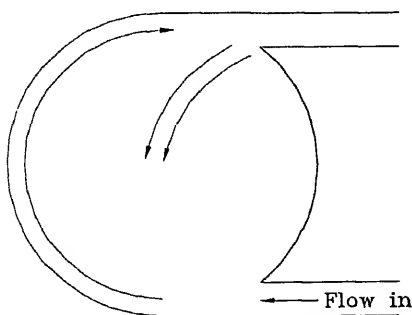
Figure 35.   Cavity did not form in the first model.

(a)   (b)
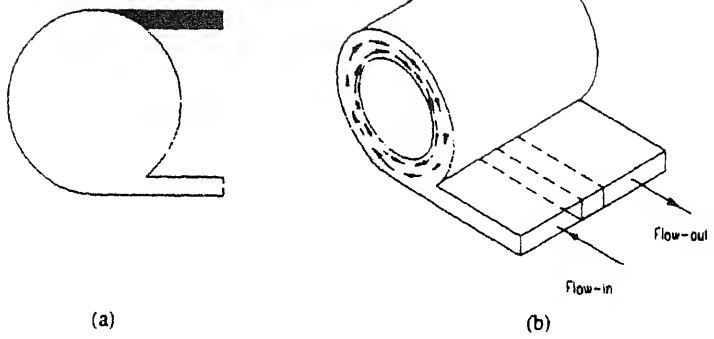
Figure 36. Blocked upper channel and divided lower one.
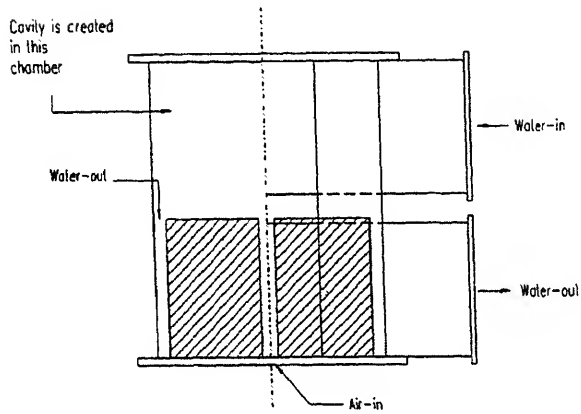
Figure 37. Our modified model.
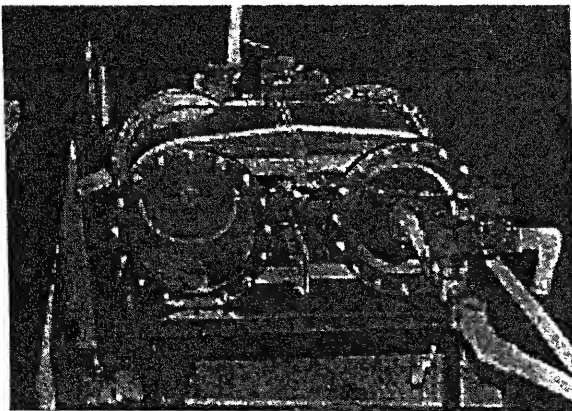
Figure 38. Experimental configuration

Figure 39. Minimum cavity of the periodic working engine.

For simplicity of pipe connections, we used a horizontal configuration for the first experiment, which caused an unstable cavity and air–water mixing. Therefore, the next design used a vertical configuration.

6. Engine redesign and experimental analysis

In order to avoid air/water mixing which occurred in the first model, we decided to adopt a vertical configuration with swirlers (figure 41). From the previous model, we realized that the most important design goal is to make the system as axisymmetric as possible, because any non-axisymmetry will most likely cause the cavity to become unstable or to collapse on the sudden application of pressure.

In order to have an axisymmetric inlet, we designed swirlers as shown in figure 42. This swirler forms evenly distributed water jets which helps to create a more stable cavity.

The number of blades on the swirler is not arbitrary and is related to the dynamics of the system. Its relationship with hydraulic load ζ and frequency f is shown in figures 43 and 44 respectively.

If we want this model to work at a modest frequency, say 5 Hz, just to demonstrate the principles, from figure 44 we find the corresponding n is 12.

The geometry of the swirler is also related to the dynamics of the system and has been carefully designed. Figure 45 shows a typical entrance geometry where angle BAD forms a
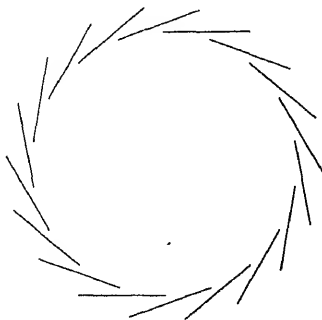
Figure 40. Water distributor.
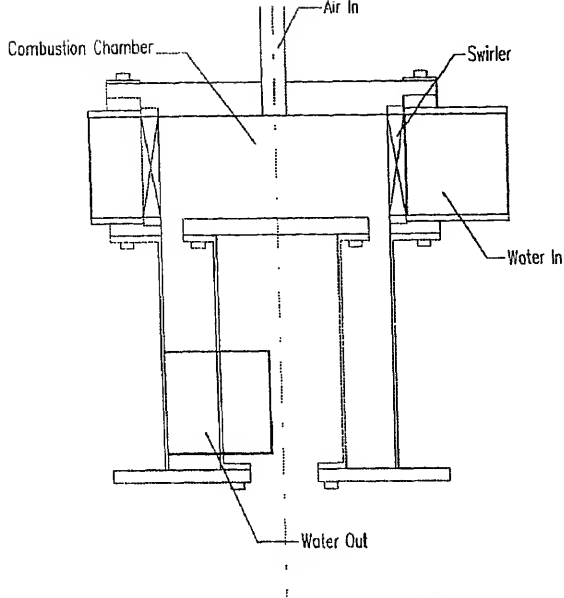
Figure 41. Vertical configuration.
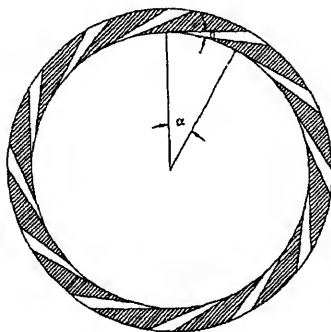
Figure 42. Swirler.
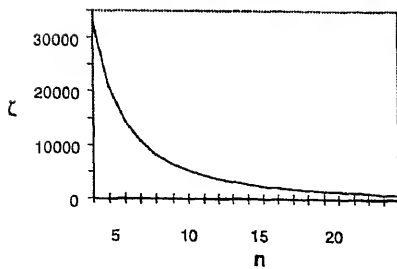
Figure 43. Load coefficient ζ as a function of n.
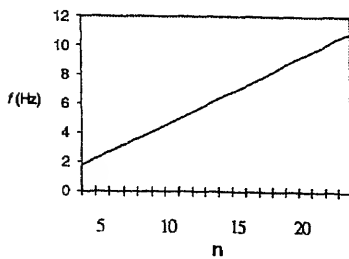
Figure 44. Oscillating frequency as a function of n.
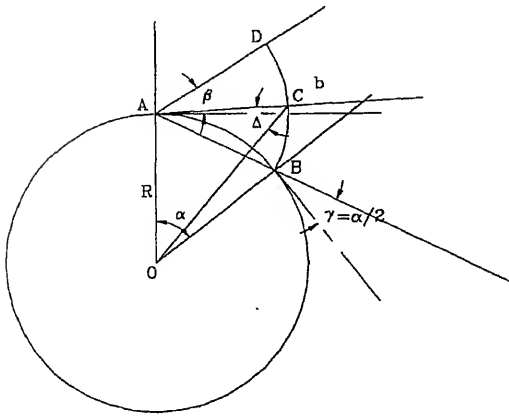
Figure 45. Swirler geometry.

single entrance. It has been proven that when the two blades intersect at point A, circulation is maximum for a given flow rate. As shown before, we introduced a new parameter $\theta = v_r/v_\varphi = |\dot{y}|$. From figure 45, it can be found that $\theta = \cot\Delta = (1 - \sin^2 \Delta)^{1/2}/\sin \Delta$. When $\beta = 0$ (tangential inlet), we have

$$\theta = \left(4 \sin^2 \frac{\alpha}{2} + 1 - 4 \sin \frac{\alpha}{2} \sin \frac{\alpha}{4} - \cos^2 \frac{\alpha}{4}\right)^{1/2} \bigg/ \cos \frac{\alpha}{4} \tag{68}$$

where $\alpha = 2\pi/n$.

We also improved the control system. First, we removed the external oscillator and used the signals from engine check valves as control parameters. Then we replaced the mechanical relay control system with a new microcontroller and solid-state relays. In order to achieve the semi-Otto cycle, air supply valves A_1 and A_2 (figure 46) must be open for only a very short time 't' (to mimic combustion), and this time should be adjustable. A_i, E_i, and I_i are the air supply, exhaust and inlet check valve for chamber i respectively.

Figure 47 shows the control sequence for the valves having a very short adjustable time t for the air inlet valves to simulate combustion.

We use a Basicon MC-1Z microcontroller to implement this control sequence. It is a fully self-contained general purpose programmable controller with CPU, RAM, ROM, real-time-clock, I/O, communications circuitry and a resident BASIC language.

Figure 48 shows the setup of the new model and figure 49 is a close-up to reveal the cavity.
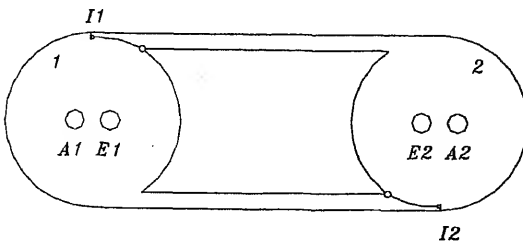
Figure 46. Schematic of the engine control.
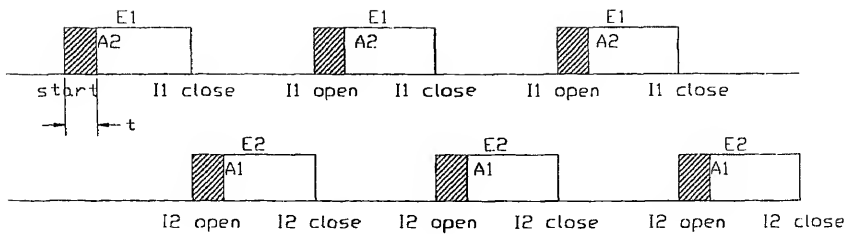
Figure 47. Timing diagram.

7. Summary

Here we briefly review the basic principles and potential advantages of the VLPE. The two cylinders are interconnected by two tangential, unidirectional flow channels containing hydromotors. Because of the tangential inlets, the liquid is caused to rotate in each cylinder and creates a vortical liquid body with a cylindrical cavity in the centre. Each cylinder has an inlet valve for the intake of fuel and air, as well as an associated exhaust valve. The cylinders can either have an electric spark plug or work in a diesel mode with fuel injectors.

The rotating liquid in the engine simultaneously performs three functions: (a) it creates a stable cavity; (b) it stores rotating kinetic energy (a liquid "flywheel"); and (c) it has effective elastic properties for damping detonation and permits a higher compression ratio than that of the conventional engines.

Our mathematical model describes the behaviour of this two-cylinder engine. Here we present the design and operating parameters of a 100 kW engine. The diameter of each cylinder is 0.1 m, the cross-sectional area of the tangential inlets and outlets is 20 cm^2 and the velocity of the input liquid is 10 m/s. The corresponding pressure drop across a hydromotor is 14 atm. The thermal efficiency for the Otto cycle is 0.6 and the compression ratio is about 10. The stability criterion for this example is 72, which ensures a large margin of stability. The acceleration in the maximum compression stage is very large
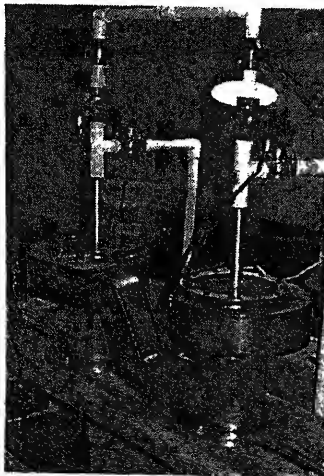
Figure 48. Setup of the engine.
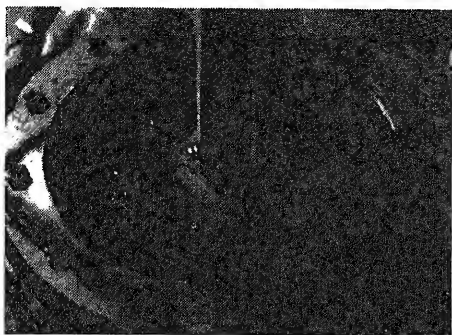
Figure 49.   Cavity.

approximately 1300 g. When running in diesel mode, the power and efficiency of an ngine of the same size will increase significantly because of higher compression ratio. he preliminary experiments have demonstrated the stable cavity in the "cold" model of rtex liquid piston engine.

The engine employing a rotating liquid as a piston has the following advantages:

) It is small, lightweight and generates large amounts of power;

) It is simple to service and repair, having low maintenance;

) The damping of detonation allows high compression ratios, and thus significantly improving thermal efficiency; emission of environmentally damaging pollutants is dramatically reduced;

) If a receiver is available to act as an accumulator, a new type of hydrodynamic braking is possible; part of the kinetic energy might be transferred to the receiver by the hydromotor acting as a hydropump to provide energy to the receiver.

Other vortex machines – concepts and applications

/e are using the features of vortex flows outlined in the introduction to develop several rtex machines, briefly mentioned below.

rtex thruster (VT):   It is essentially a bladeless helicopter, a real 'flying saucer' of ractical relevance. The basic working principle of this device is to create a strong swirling ow which produces a low pressure above a surface, thus generating lift or thrust. The virl in the incoming flow is achieved via an open vortex chamber, and the resulting low ressure rarefaction is intensified by an airfoil-shaped diffuser, which ensures an attached ow without separation. Referring to figure 50, the propulsion device has an open vortex namber (5) which serves to create a strong swirling open fluid jet; (1) is a source of ressurized fluid, and (2) is a header for uniform flow distribution. The flow enters the virler (3), acquires strong angular momentum and exits over the diffuser (4), where it verses direction (by the Coanda effect) and is ejected into the ambient fluid. Such a ow is like an artificial tornado, creating a strong rarefaction (low-pressure) zone on the
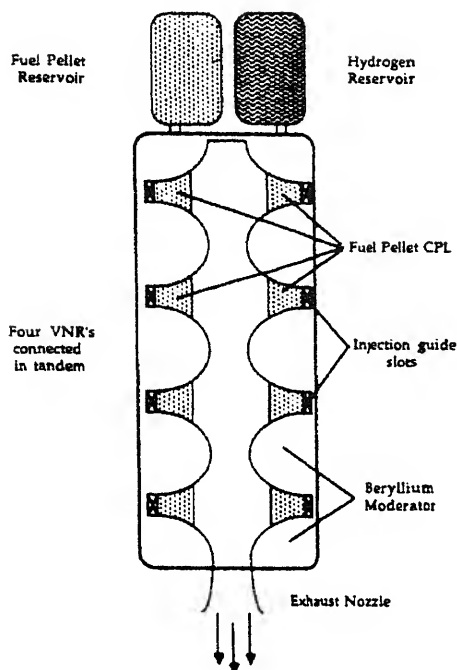
Figure 53.   Schematic of VNR.

scroll. This allows uniform distribution of the fluid along the guide device (2). The fluid then enters the vortex chamber with appropriately profiled walls (4) and leaves through exits (6). The fluid plays a two-fold role: First, it removes heat as it flows through the particle layer; second, rotation of the fluid ensures confinement of the layer.

The high centrifugal acceleration makes the VCR a more efficient fluidized bed than conventional technology which utilizes gravity. VCR is particularly well-suited in chemical reactors using catalysts, and in burning pure, granular aluminum as fuel for military and aerospace applications. The swirling flow in the vortex chamber with the exhaust at the centre guarantees residence of the particles (catalysts, reacting fuel, etc.) within the chamber while they are essentially consumed, thus significantly increasing efficiency and reducing wastage. The fluidized layer would also be useful in the agriculture industry for efficient surface drying and separation of grain.

A specific potential application of this technology is a vortex nuclear reactor (VNR), which employs a concentrated pseudofluidized layer (CPL). The CPL consists of small spherical fuel pellets rotating in a special vortex chamber surrounded by a fluid (liquid or gas) acting as a heat transfer medium. To illustrate its basic operation, the active section of the VNR for space applications is shown schematically in figure 53. It consists of granular fissionable material formed into a CPL in a special vortex chamber.

The VNR consists of several parts: fuel pellet and hydrogen reservoirs, fuel pellet CPL, injection guide slots, beryllium moderator, and exhaust nozzle. It has the following possible uses:

(i) nuclear reactor with high specific power output utilizing radically reduced amounts

(iii) high neutron flux research reactor and breeder.

The VNR has significant safety advantages over conventional nuclear reactors because of minimal amount of fuel pellets in residence in the chamber at any time, very high specific power, compact size for ease of maneuverability, and a self-cleaning capability for encrusted fuel elements. This latter feature permits use of the VNR with heat transfer media such as oil or sea water.

Bubbling centrifuge (BC):   Vortex bubbling devices are used to enhance heat and mass transfer processes between liquids and gases by injecting the gas into the rotating, centrifugally stabilized liquid. Direct control of the strength of the centrifugal force permits the construction of very compact machines with high efficiency compared with conventional devices which rely on gravitational force for their operation. Since this force cannot be varied, such an apparatus must frequently be very large in order to process the necessary volume of gas. Thus, there are major expenses associated with the construction, transportation, and operation of these devices. High maintenance costs are also associated with the frequent removal of interior scale and deposits. With its drastically reduced size and simplicity of operation, vortex barbotage machinery offers significant reductions in costs, at all levels. In the condensed stable state (of the bistable states) of a gas–liquid mixture in a vortex chamber (figure 54a), a self-organized, rotating homogeneous layer of foam is formed near the periphery of the chamber (Goldshtik 1981). In this layer, the high centrifugal acceleration suppresses the entrainment of drops and increases the liquid–gas interface; this provides a very large reaction surface area per unit volume and enhances interfacial transport, further accentuated by turbulent mixing.

Principle of operation – Gas is injected from the walls into a swirling flow of liquid, forming a rotating bubbly or frothy layer. Extensive theoretical and experimental studies have determined the precise vortex chamber shape and other design parameters necessary to ensure a stable froth layer.
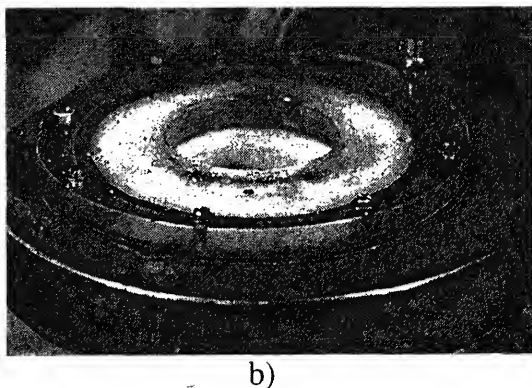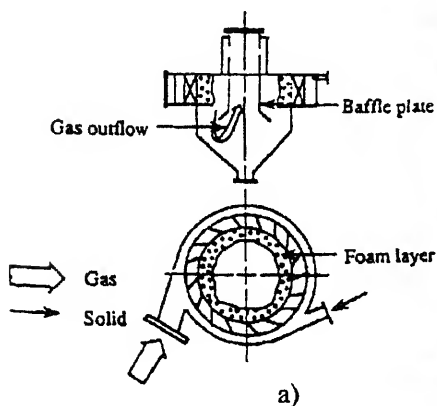
Figure 54.   Bubbling device and layer.

Under these conditions, the stable froth layer maintains a liquid–gas force balance at the chamber walls as well as in the interior, ensuring uniformity in gas/liquid interaction as the gas passes through the liquid. A photograph of such a layer in a recent laboratory experiment is shown in figure 54b. The tangential injection of gas at the chamber walls through small nozzles creates bubble streams which then transfer their angular momentum to the liquid. Under normal operation, this generates centrifugal forces in the chamber which are 10 to 100 times greater than gravity. Such increased forces permit the use of much smaller diameter bubbles, which greatly increases the gas/liquid interfacial contact area. In addition, the very high speeds of the gas bubbles through the liquid and the strong secondary cross-flows in the liquid significantly enhance the liquid/gas interfacial interaction. Experiments have shown that a homogeneous froth layer can be maintained with correct design, without disruption by bubble coalescence or other nonuniformities. Due to these factors, a vortex barbotage apparatus can be built that can process the same volume of gas as a conventional device which is ten times its physical size and which consumes more energy. Here is a summary of the advantages of the vortex bubbling machine.

(i) Compared to conventional bubbling equipment, it is up to 10 times smaller and weighs as little as 1000 times less than comparable gas processing devices.

(ii) It has very high efficiency due to enhanced thermal and mass transfer processes.

(iii) Its small size and weight make transportation, and possible installation in existing facilities easy.

It can function with a wide range of gas and liquid flow ratios; thus the device can operate under strong inertial loading (e.g. on a moving vehicle), or in a micro-gravity environment (space applications); and works with the froth layer detached from the chamber walls; this reduces formation of scale and encrustation on walls, reducing maintenance costs.

Possible applications include:

(i) both particulate and vapour pollutant emission control at fossil fuel power generating plants. The worldwide market for such devices is in billions of dollars;

(ii) absorption and desorption of gases: e.g. carbon dioxide, nitrous oxide, oxygenation and carbonization of solutions, and many other chemical technological processes;

(iii) vaporization and condensation of liquids, such as in the thickening, heating and cooling of solutions, e.g. milk;

(iv) petroleum refining having a wide variety of applications;

(v) three-phase processes, in which the solid particle/liquid mixture is kept in a gas stream; this has important applications in the mining industry, e.g. flotation or segregation of fine gold from mine waste, and
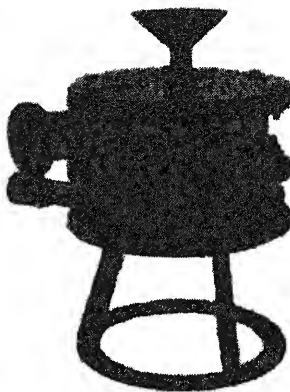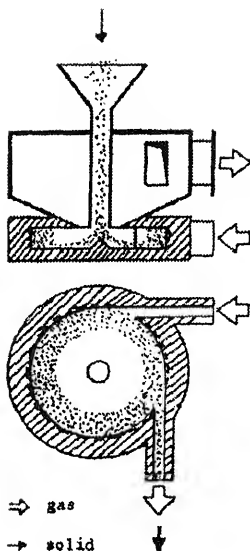
Figure 55. Vortex mill.

(vi) vortex nuclear reactor with a liquid uranium active zone. Because the critical temperature of liquid uranium is $\sim 20{,}000\,\mathrm{K}$, it is possible to heat the hydrogen working fluid up to temperatures of 6,000–7,000 K, giving a specific impulse of $\sim 1{,}500\,\mathrm{s}$.

Vortex mill (VM):   Theory and experiments (Goldshtik 1981, 1985) show that large particles can be ground efficiently while in the rarefied stable state (of the bistable states) of a solid–gas mixture in a rotating flow (figure 4). Adding acoustic resonators to the vortex chamber (figure 55) produces an intense sound field which considerably improves the grinding performance. We conjecture that this sound aids in grinding by preventing the healing of collision cracks between successive impacts, thereby facilitating crack propagation. Operating parameters can be modified appropriately to allow extremely fine grinding, and thus VM has a very crucial use, e.g. in obtaining pure (99.9999%) yttrium powder needed for making sintered film coatings for cathodes in the electronics industry. This is also an important ingredient in the manufacture of high-temperature superconductors. VM can also be used in cement production, and in the mixing of cement and fly ash in the manufacture of concrete slabs. In power plants, VM can permit on-site grinding of coal immediately prior to combustion, increasing combustion efficiency and eliminating off-site pulverizing to reduce transportation costs; in fact, it can be an integrated part of the combustor. VM can grind plastic as well as brittle material and thus can help grind used tyres for recycling, and polymers for use as additives for drag reduction in pipelines, submarines and boats etc. It can also be used as a very efficient liquid atomizer.

9.   List of symbols

c_p, c_v	specific heats;
E	temperature ratio of combustion;
f	nondimensional pressure; frequency;
h	chamber height;
k	adiabatic exponent, c_p/c_v;
n	number of slits on a swirler;
n_0	stoichiometric air/fuel ratio;
p	pressure; p_0 – lowest pressure in a cycle;
q_f	enthalpy of reaction;
Q	area flow rate; Q_v – volume flow rate;
r	oscillating cavity radius;
R	universal gas constant; cylinder radius;
S	entropy;
S_t	stability coefficient;
T	temperature and torque;
V	volume and inlet velocity; V_0 – volume at pressure p_0;
v_r, v_φ	radial and tangential velocities of interface;
W	engine power;
y	nondimensional volume; y_0 – initial value of y;
α	swirler distribution angle, relative air/fuel ratio;
Γ	circulation;
ϵ	compression ratio;
η	thermal efficiency;
ζ	load coefficient;
ν	kinematic viscosity;
Π	pressure parameter;
θ	velocity ratio;
ρ	density;
σ	ratio of the total gas volume to total cylinder volume;
τ	nondimensional time scale;
τ_c	period of oscillation;
χ	ratio of maximum to minimum pressures during one cycle;
ω	eigenfrequency; ω_0 – eigenfrequency for small oscillation.

References

Batchelor G K 1967 *An introduction to fluid dynamics* (Cambridge: Cambridge University P

Faires, Virgil M 1970 *Thermodynamics* (New York: MacMillan)

Goldshtik M A 1981a *Vortex flows* (Nauka: USSR Academy of Sciences)

Goldshtik M A 1981b Heat and mass transfer in a twisted gas-liquid layer. *J. Appl. Mech. Phys.* 22: 850–856

Goldshtik M A 1984 *Transfer processes in granular layers* (Nauka: USSR Academy of Scie

Goldshtik M A 1985 Variational model of a turbulent rotating flow. *Fluid Dyn.* 20: 353–36

Goldshtik M A 1992 Engine employing rotating liquid as a piston. *US Patent*, No. 5,127,369

Greenspan H P 1969 *The theory of rotating flows* (Cambridge: University Press)

Gupta A K, Lilley D G, Syred N 1984 *Swirl flows* (Tunbridge Wells, UK: Abacus Press)

Hussain F 1986 Coherent structures and turbulence. *J. Fluid Mech.* 173: 303–356

Popular mechanics 1995 The incredible shrinking engine. January, p. 24

Shtern V, Hussain F 1995 Hysteresis in swirling jets. *J. Fluid Mech.* 309: 1–44

West C D 1983 *Liquid piston stirling engines* (New York: Van Nostrand Reinhold)

Jet flames from noncircular burners

S R GOLLAHALLI

School of Aerospace and Mechanical Engineering, The University of Oklahoma, Norman, OK 73019, USA
e-mail: gollahal@lincoln.ecn.ou.edu; gollahal@ou.edu

Abstract. Gas jets from noncircular exits entrain more air from surroundings than jets from circular exits of equivalent area. Because the mixing rate of fuel and air governs the combustion and pollutant emission of diffusion flames and partially premixed flames, noncircular geometries offer a passive control of the flame characteristics. In this paper, the literature on nonburning, noncircular jets is reviewed and recent studies on noncircular jet flames are discussed with focus on the work conducted in the author's laboratory.

Keywords. Gas jets; noncircular burners; jet flames; flame characteristics.

1. Introduction

Combustion and pollutant characteristics of fuel gas jets are controlled by the mixing rates of fuel and ambient air. Hence, manipulation of fluid mechanics in the vicinity of the burner mouth by varying the exit geometry appears to be an attractive method of controlling the thermochemical processes to increase combustion efficiency and reduce pollutant emission from diffusion flames. This paper presents the research on this topic with focus on the work at the University of Oklahoma.

2. Background

Noncircular exits are often considered to be passive and inexpensive methods of controlling the characteristics of isothermal jets. The geometries usually considered are elliptic, triangular, and rectangular with a small aspect ratio of 2 to 4. The nonaxisymmetric flow field tends to remove its instability by gradually becoming symmetric downstream. In the process, the smaller dimension tends to become larger and the larger dimension tends to become smaller, leading to a phenomenon called axis-switching. Before achieving symmetry, a few switchings usually occur. This characteristic of noncircular jets has been well documented (Sforza *et al* 1966; Schadow *et al* 1984; Gutmark *et al* 1985; Ho & Gutmark 1987; Hussain & Hussain 1989; Quinn 1989; Gollahalli *et al* 1992). The locations and the number of switch-overs are strongly dependent on the initial conditions, the aspect

ratio of the nozzle exit, and when excited on the flow Strouhal number. The dominance of coherent structures and azimuthal instabilities are stated as responsible for axis-switching. Koshigoe *et al* (1988) have theoretically analysed the condition for deformation of noncircular jets. The analysis of large eccentricity elliptic jets performed by Crighton (1973) has shown them to be spatially as well as temporally stable. Ho (1986) states that entrainment in two-dimensional shear layers and axisymmetric jets occurs during the vortex-merging event. In nonaxisymmetric jets, however, vortex-merging and azimuthal deformation occur simultaneously.

In comparison to the number of studies on nonreacting jets, work on the combustion of noncircular jets is severely limited. Schadow *et al* (1984) have found in a large-scale study that the combustion efficiency in elliptic jets is about 10% higher than in circular jets. Better mixing is shown to be the cause of this improvement. Gollahalli and his associates (Schadow *et al* 1984; Gollahalli & Prabhu 1990; Prabhu & Gollahalli 1990; Kamal & Gollahalli 1993; Subba & Gollahalli 1995) have investigated the structure and pollutant formation in flames of noncircular jets at various conditions. Cannon & Queiroz (1994), have studied the time-resolved temperature measurements in an elliptic turbulent diffusion flame. Recently, Papanikolau & Wierzba (1995, 1996) have investigated the lift-off and blow-out characteristics of diffusion flames. These studies are discussed in the following.

3. Turbulent diffusion flames

The results in this section are extracted primarily from the works of Prabhu (1989), Khanna (1990), and Kamal (1995). The experiments by Prabhu and Khanna were carried out on the same burner and experimental apparatus. Prabhu investigated the thermochemical characteristics and Khanna studied fluid mechanics of the turbulent gas diffusion flames from elliptic nozzles. Kamal recently focused his work on the coupling of burner size and exit Reynolds number on the noncircular burner effects.

3.1 *Experimental apparatus and instrumentation*

The experiments in these studies were performed in a vertical flow combustion apparatus. This facility had a combustion chamber of square cross-section (76 cm × 76 cm) and 117 cm height. The chamber had air-cooled Pyrex glass windows (20 cm × 20 cm × 92 cm) on all the side walls. While drawing gas samples and probing the temperature field, one of the glass windows was replaced with a slotted metal sheet through which the probes were introduced. The glass windows extended down to the floor level of the chamber to permit optical probing of the flames in the vicinity of the burner mouth. The burner assembly mounted to the floor of the chamber consisted of a contoured nozzle section located concentrically inside a cylindrical chamber. The fuel nozzle had four parts: a divergent inlet section, a cylindrical settling section, a convergent contoured section, and interchangeable nozzle tips. Nozzle tips of different configuration could be attached to the contraction section. Except in Kamal's work on burner size effects, a circular nozzle tip of exit diameter 9.5 mm i.d. was used. The elliptic nozzle tips (major diameter = 16.4 mm and minor diameter = 5.5 mm) had the same exit area as that of the circular nozzle. The
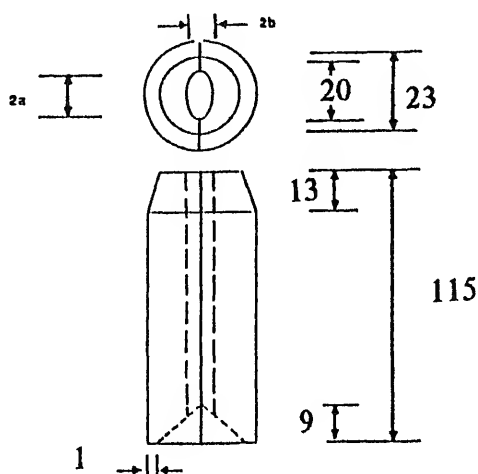
Figure 1. Details of the elliptic nozzle tips for diffusion flame studies.

ratio of the length of the elliptic cylindrical portion to the major diameter of these nozzle tips was at least 7, ensuring the ellipticity of the initial geometry of the fuel jet. Figure 1 shows the details of the elliptic nozzle tips. A mixture of propane and nitrogen was used as the jet fluid for the reason discussed later in relation to flame stability results. Combustion air was supplied from an oil-less rotary-vane compressor to the annulus surrounding the nozzle. An assembly of a honeycomb section and two sets of fine-mesh screens ensured uniform and low-turbulence air flow concentric with the fuel jet. In all experiments reported in this study the velocity of co-flow air was maintained at 0.66 m/s. The fuel jets were ignited with a pilot Bunsen flame which was withdrawn during tests.

Direct colour photography with an exposure time of 1 second was used to determine the shape and dimensions of the visible flame. A two-mirror (200 mm diameter) Z-arrangement schlieren system with a xenon stroboscopic light source of flash duration 1.5 μs was used to visualize the flow structure in the near-nozzle region of the flames (Khanna 1990). The temperature field was probed with a thermocouple. The composition field was determined by gas sampling and analytical instruments consisting of nondispersive infrared analysers, a polarographic analyser, a chemiluminescent analyser, and a gas chromatograph. Soot concentration was determined by filtering known volumes of diluted gas samples. Radiation emitted from flames was measured with a wide view angle radiometer. The flow field in the vicinity of the flame base was probed with a laser Doppler velocimeter. Both nozzle and air-flow streams were seeded with magnesium oxide particles. A microcomputer-based data acquisition system was used to acquire and process the output data of the LDV. The details of instrumentation and error analysis are given by Prabhu (1989), Khanna (1990), and Kamal (1995). Table 1 shows the nominal experimental conditions.

3.2 *Flame stability*

Most of the studies on flame life-off have focused on the burner exit velocity as the

Table 1. Nominal experimental conditions.

Type of flame	Turbulent diffusion flame	Partially-premixed flame
Fuel	Propane and nitrogen mixture	Natural gas
HHV (MJ/kg)	11–18	37.3
Energy input rate (kW)	750–1232	5.86
Fuel jet velocity (m/s)	—	49.1
Fuel jet Re_j	—	6500
Burner exit velocity (m/s)	5.60	1.275
Burner port Re_p	7810	3640
Ambient temperature (K)	297	295
Ambient pressure (mm Hg)	745	749
Ambient relative humidity	65%	65(%)

delineated. Prasad *et al* (1991) employed a method in which the lift-off phenomenon can be studied in terms of the reactant concentration in the jet fluid while keeping the burner exit velocity constant. This method is useful to isolate the effects of parameters like nozzle shape by decoupling their influences from the accompanying changes through the jet exit velocity. Hence, a mixture of N_2 and C_3H_8 was used as the jet fluid and the concentration of C_3H_8 at transitions was used as an indicator of the flame stability.

Figure 2 compares the transition characteristics of flames from circular and elliptic (aspect ratio 3 : 1) nozzles when the mole fraction of propane, x_p, in the jet stream is decreased while keeping the exit velocity approximately constant at 5.65 ± 0.30 m/s. The nominal exit velocity corresponds to a Reynolds number of 4740 in the case of the circular nozzle, and 7810 and 2620, based on the major and minor diameters in the case of the elliptic nozzle. Here L is the flame standoff distance and D_{eq} is the diameter of the circular
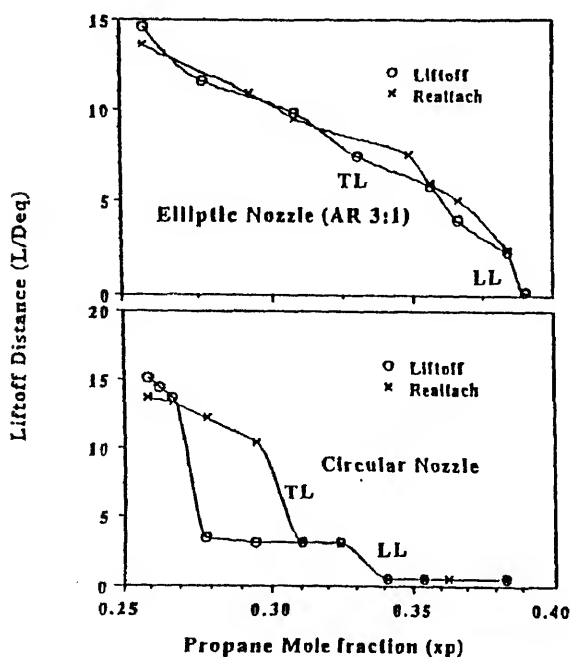
Figure 2. Lift-off and reattachment characteristics of turbulent diffusion flames from circular and elliptic nozzles.

ozzle of equal exit area. When x_p in the jet stream of the circular nozzle was above .34 the flame was attached to the burner. At x_p equal to 0.32, the flame base lifted off the urner and stabilized about 3 nozzle diameters above. When x_p was between 0.32 and 0.34 ne flame base position was not well-defined and the flame was seen to be detached from nly a part of the nozzle rim. The flame edges appear smooth and the flow near the flame ase appears laminar. This transition is denoted as "Laminar Liftoff" and is shown as LL n the figure. When propane mole fraction is further decreased to 0.28, the liftoff height uddenly increases to about 10 diameters while the flame base widens to about 4 nozzle iameters. Flow near the flame base is turbulent in this configuration. This transition is enoted as "Turbulent Liftoff" and is indicated as TL on the figures. When x_p is decreased urther L increases slowly before flame blows out at $x_p = 0.24$. With the elliptic nozzle, ne first transition LL occurs at $x_p = 0.38$, the second transition TL occurs at $x_p = 0.33$, nd flame blows out at $x_p = 0.24$. This shows that elliptic nozzle flames require more eactive component in the jet fluid to avoid transitions of burner-attached flame to liftoff nd blowout conditions than circular nozzle flames. Also, the flame-standoff distance does ot remain constant between the first and second transitions in the elliptic nozzle flames, n contrast to the circular nozzle flames.

When propane mole fraction in the jet fluid is increased while the flame from the ircular nozzle is in the lifted condition, the L/D_{eq} versus x_p curve does not retrace its ath and exhibits hysteresis similar to the hysteresis noticed in earlier studies (Scholfield ∠ Garside 1949; Gollahalli *et al* 1987). In those studies, the differences in the turbulence haracteristics of the flame-standoff region upstream of the flame base between the attached nd lifted flames were thought to be responsible for the hysteresis. The hysteresis behaviour, owever, is not significant in the flame from the elliptic nozzle. The earlier initiation of 1stabilities and consequently higher turbulence level in the elliptic nozzle jets seem to be esponsible for the negligible hysteresis in the liftoff-reattachment characteristics of the lliptic nozzle flame.

.3 *Fuel concentration profiles*

he axial concentration profiles of propane in the attached flames from circular and elliptic ozzles are shown in figure 3. We notice that fuel concentration decreases more slowly long the axis and the fuel persists longer in the circular nozzle flame than in the elliptic
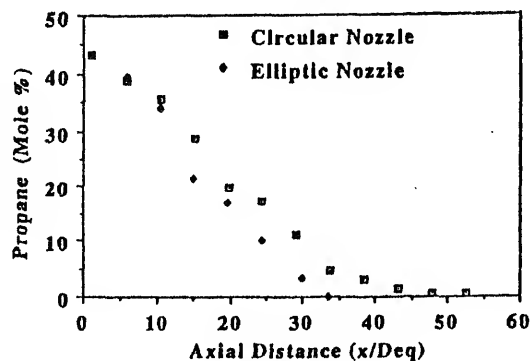
Figure 3. Effect of nozzle shape on the axial concentration profiles of fuel in turbulent diffusion flames from circular and elliptic nozzles.

nozzle flame. For instance, the value of x_p decreases to 0.005 at $x/D = 48$ in the circular nozzle flame, whereas the same value is reached at $x/D_{eq} = 34$ in the elliptic nozzle flame. The rapid depletion of propane is caused by the greater dilution due to the higher entrainment in the elliptic nozzle flame.

3.4 Temperature profiles

Figure 4 shows the radial temperature profiles at three axial levels in the circular and elliptical nozzle flames attached to the burner. These locations are in the near-nozzle, midflame, and far-nozzle regions of the flames. In both flames the near-nozzle profile has a sharp off-axis peak. A comparison of these profiles reveals that the circular nozzle flame has a slightly higher peak temperature compared to the elliptic nozzle flame in the near-nozzle region. However, in the midflame and far-nozzle regions, the elliptic nozzle flame has higher peak values. As homogeneous gas phase reactions are dominant in the near-nozzle region, the more rapid development of a shear layer and the higher degree of mixing with air in the elliptic nozzle flame could lead to a lower peak temperature. On the other hand, in the midflame and far-nozzle regions, the kinetics-controlled heterogeneous soot oxidation reactions are the primary mechanisms of heat release (Gollahalli 1977). These kinetics-controlled processes are enhanced in the elliptic nozzle flame by the larger oxygen availability, thereby leading to higher rates of burning of soot. The higher heat release rates and the lower heat loss rates due to lower concentrations of soot result in higher temperatures in the midflame and far-nozzle regions of the elliptic nozzle flame, compared to the temperatures in the corresponding regions of the circular nozzle flame. It appears that the dilution effect of increased entrainment of air into the flame is overshadowed by the enhancement of soot combustion which results in higher temperature in the midflame and far-nozzle regions of the elliptical nozzle flames. The differences in the jet growth rate determined from velocity measurements in circular and elliptic nozzle flames presented later are in conformity with the temperature profiles.

3.5 Nitric oxide and particulate concentration profiles

Figure 5 shows the radial concentration profiles of nitric oxide at three axial levels in the attached flames from the circular and elliptic nozzles at the same conditions at which the temperature field was probed. Except in the far-nozzle regions, the differences in the peak concentration of nitric oxide in the circular and elliptic nozzle flames follow the trends of temperature profiles discussed above, as expected of thermal nitric oxide produced through Zeldovich reactions. In the near-nozzle region, the peak concentration of nitric oxide is higher in the circular nozzle flame, because of the thinner interfacial reaction zone that results in a higher peak temperature than in the near-nozzle region of the elliptic nozzle flame. In the midflame region, the elliptic nozzle flame has a higher nitric oxide concentration than the circular nozzle flame, which also follows the trend of temperature. In the far-nozzle region, the average of the NO concentrations on the major and minor axes of the elliptic nozzle is slightly less than the NO concentration level in the circular nozzle flame.
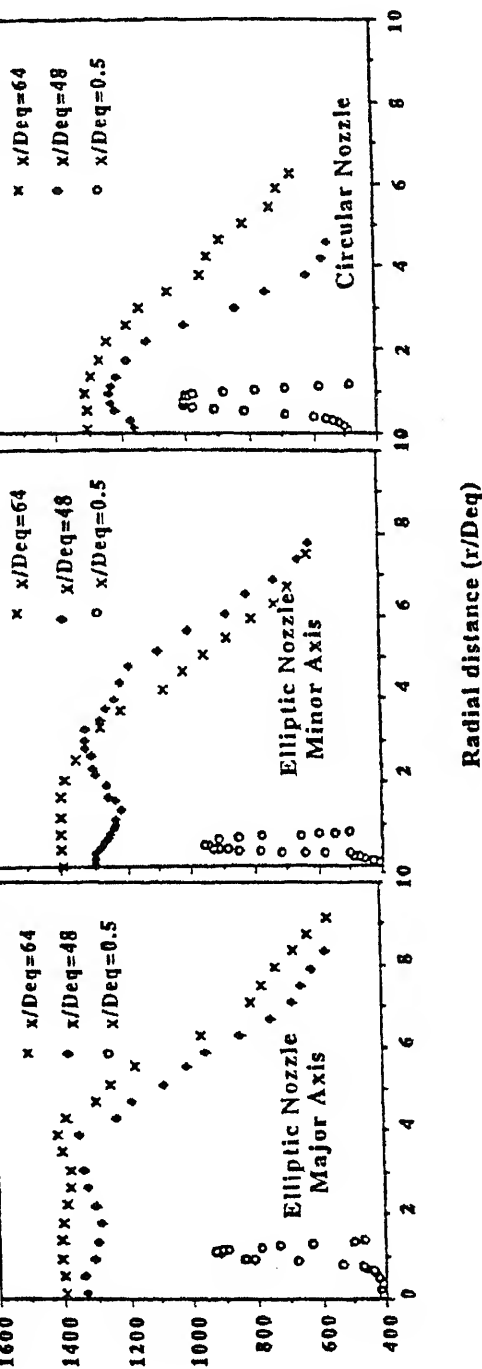
Figure 4. Radial temperature profiles in turbulent diffusion flames from circular and elliptic nozzles.
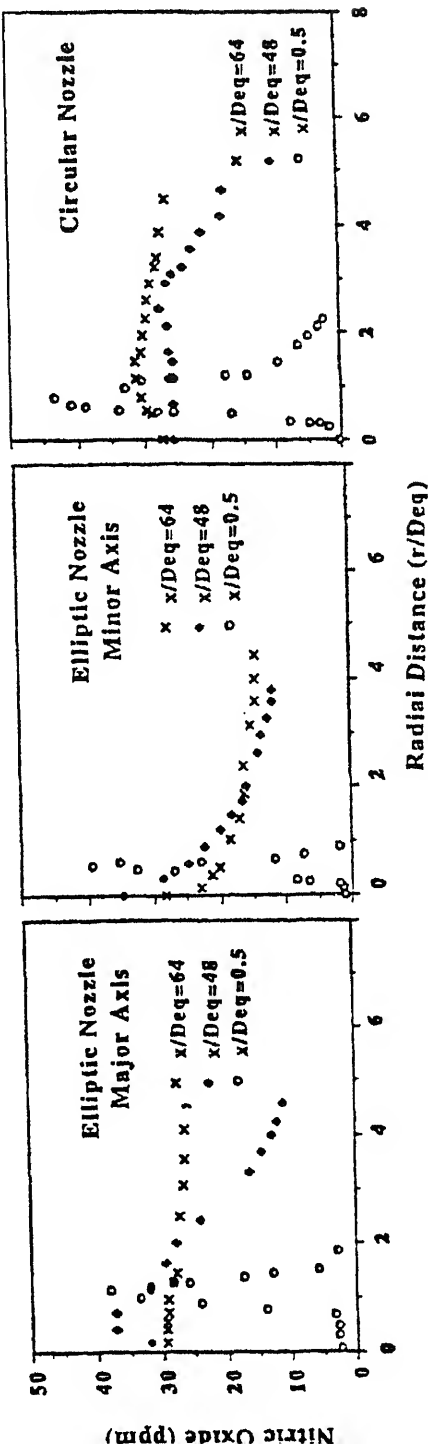
Figure 5. Radial nitric oxide concentration profiles in turbulent diffusion flames from circular and elliptic nozzles.
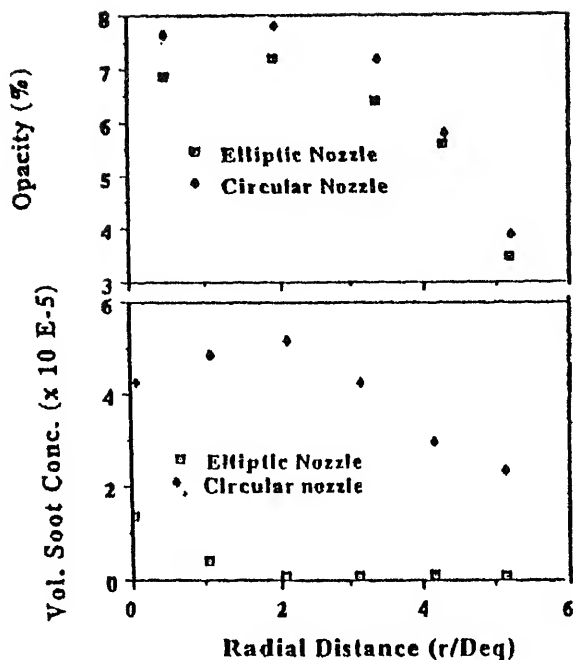
Figure 6. Effects of nozzle shape on soot concentration and opacity in turbulent diffusion flames from circular and elliptic nozzles.

The radial variations of particulate concentration and opacity of the combustion product samples drawn from the far-nozzle regions of the circular and elliptic nozzle flames are shown in figure 6. Clearly, soot concentration in the elliptic flame is smaller than in the circular flame. The difference in the opacities of samples from the two flames is not as large as the difference in the mass concentration of soot, probably because of the variance in the size distribution of the particulates. The lower soot concentrations in the elliptic nozzle flame are in conformity with the effects expected from the higher entrainment of air. The higher entrainment of air (Gutmark *et al* 1985) promotes soot oxidation rates in the far-nozzle region. Consequently, the particulate concentration in the elliptic nozzle flame is lower.

3.6   *Nitrogen oxides emission index*

Figure 7 shows the comparison of the emission index (EI) of NO_x of the circular and elliptic nozzle flames. The values of emission index (ng/J) of elliptic nozzles are normalized with the emission index of circular nozzle. The letter E corresponds to the elliptic nozzle and the number following the letter E refers to its aspect ratio. It is clear from the figure that in turbulent diffusion flames the $EINO_x$ is not significantly different for elliptic and circular nozzles. This can be traced to the opposing variations of the peak NO_x concentrations in different regions of the flame. The smaller value of $EINO_x$ reflects the trends of the NO concentrations in the far-nozzle region. The degree of reduction, however, varies
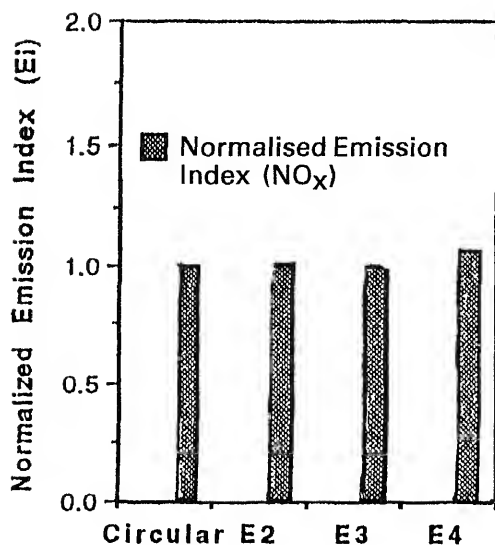
Figure 7. Comparison of nitric oxide emission index in turbulent diffusion flames from circular and elliptic nozzles.

Laminar partially premixed flames

The studies presented here on this topic are taken from the works of Kolluri (1992), Rao (1994), and Subba (1995). The natural gas furnaces of residential and commercial heating systems employ partially premixed laminar flame burners. These burners are operated with low supply pressures, typically of the order of a few centimeters of water column, and combustion occurs generally at atmospheric pressure. Hence, the designer's choice of varying the operating parameters to control fuel-air mixing rates and pollutant emission characteristics of these burners is severely limited. Motivated by the argument that air entrainment into isothermal gas jets can be significantly altered by the passive approach of using the noncircular geometries for the nozzle exit, a research program in the author's laboratory has been focused on changing the shapes of various crucial components of an inshot gas burner of a residential natural gas furnace to noncircular geometry. The exhaust emission indices of NO, NO_x, and CO, in-flame profiles of temperature and species concentrations, air entrainment rate, and flow patterns have been studied for noncircular geometries (elliptic, triangular, and rectangular) of fuel nozzles, venturis, and burner exit ports. The discussion here is limited to only *elliptic* geometries of the *burner exit port*.

.1 Burner apparatus

Figure 8 shows the sketch of an inshot burner rated at 5.46 kW used in a natural gas-fired residential furnace. Experiments were conducted in the combustion chamber and the instrumentation facilities described in § 3.1. The natural gas was supplied from the lab supply line and heating values were monitored on a daily basis. The fuel flow rate to the burner was adjusted in order to keep the energy input rate to the burner constant. The flow velocity and the area of the elliptic burners were maintained the same as those of the circular burner, and hence the observed changes in flame characteristics could be
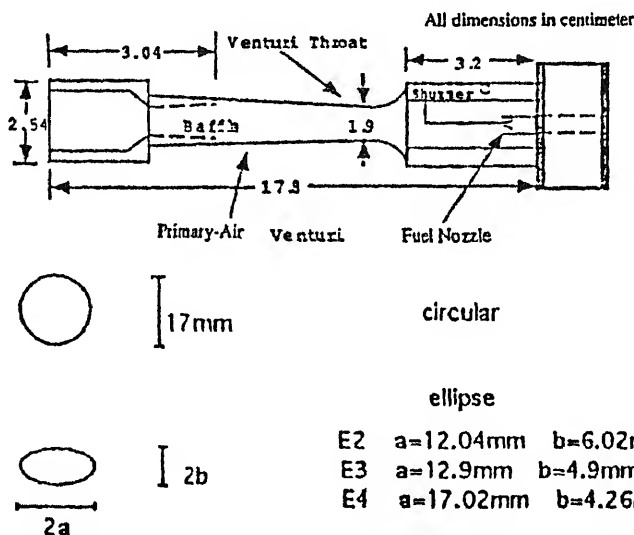
Figure 8.   Sketch of the inshot burner with standard and modified tips.

attributed to only the shape of the burner exit port. Table 1 also contains the experimental conditions in this series of studies.

4.2   Results

Figure 9 shows the effects of elliptic geometry of the burner tips on the ambient air-entrainment into the nonburning jets. Entrainment rate was determined by measuring the concentration of oxygen at 15 D_{eq} corresponding to the visible flame length. It is clear that the elliptic burner tips increase air-entrainment by as much as 30%. More interestingly, the entrainment change is not monototic with the aspect ratio of the burner tip. The shadowgraph pictures of the flames (Subba 1995) with the standard circular and noncircular burner tips show that the flow emerging from the burner port has azimuthal instabilities
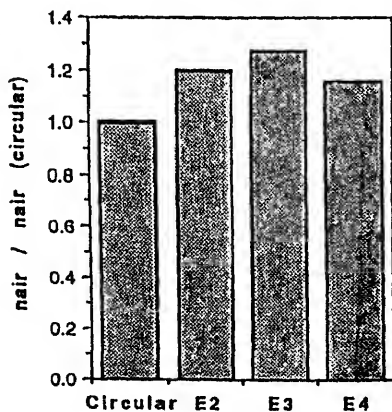
Figure 9.   Air-entrainment into the partially premixed flames of circular and elliptic tip burners.
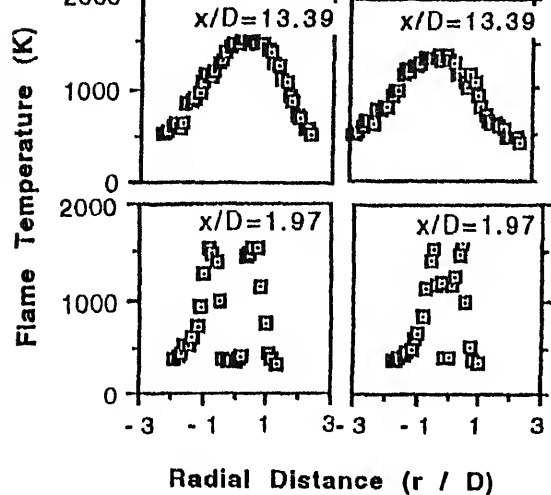
Figure 10. Comparison of temperature profiles in the partially premixed flames of circular and elliptic tip burners.

indicated by the axial streaks in both flames. However, the elliptic burner flame shows more numerous streaks confirming that the azimuthal instabilities are enhanced by the elliptic geometry of the burner.

Figure 10 shows the temperature profiles in the near- and far-nozzle regions of the flames of the circular and the elliptic burners. It is clear that the trends of these profiles are strikingly similar indicating that there is no dramatic shift in the dominant physico-chemical processes due to the shape of the burner tips. The peak values of far-nozzle temperature in the elliptic burner flames, although not conspicuous in the figure because of the compressed scale, are lower than in the circular burner flames. This decrease can be attributed to the higher air-entrainment and lower soot production in the elliptic burner flames. In the near-nozzle region, the temperatures are not significantly different in the
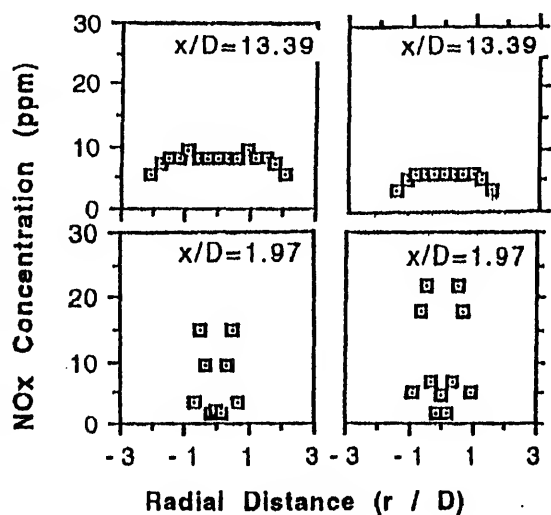
Figure 11. Comparison of nitrogen oxide concentration profiles in the partially premixed flames of circular and elliptic tip burners.
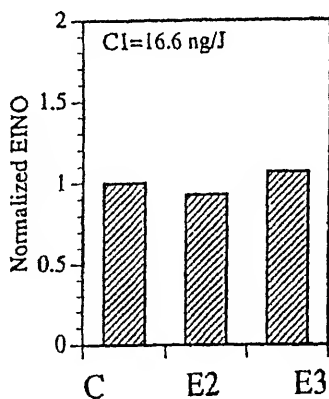
Figure 12. Nitrogen oxides emission index in the partially premixed flames of circular and elliptic tip burners.

two flames, presumably due to the fact that the increase in the oxidation rates of the fuel pyrolysis fragment species compensates for the increased dilution caused by the higher air-entrainment in that region.

Figure 11 shows the NO_x concentration profiles in the circular and the elliptic nozzle (aspect ratio 4) flames. Similar to the temperature profiles, the NO_x concentration profiles also show double hump structure in the near-nozzle and midflame regions and either flat or single axial peak structure in the far-nozzle region. It is clear that interfacial combustion at the flame edges continues to be dominant until the midflame region. In conformity with the temperature profiles and entrainment results, the values of the peak NO_x concentration in the far-nozzle region are generally lower in the elliptic burner flame.

Figure 12 shows the effect of burner shape on emission indices of NO_x. It is clear that the elliptic shape E2 has a favorable effect on the pollutant emission indices of the burner, although the effect is less than about 15%.

5. Concluding remarks

Flow visualization studies of the structure of gas jets and flames have shown that the large scale coherent structures initiated by Kelvin–Helmholtz instabilities are responsible for the entrainment of surrounding fluid (Ho & Gutmark 1987). Also, it has been shown that noncircular geometries of the nozzles introduce many azimuthal instabilities which interact with these organized structures, and consequently, produce differential radial growth rates of jet and the consequent axis switching. Thus, air-entrainment rate and combustion characteristics of flames are affected by the noncircular shape of the burner, which has been demonstrated in the case of turbulent propane *diffusion* flames. Although the *partially premixed* flames were laminar, the results show that somewhat similar features exist. Since the burner exit conditions were such that the flames were not fully turbulent, the effects on entrainment and temperature field in this case seem to be smaller than in the turbulent case. The higher air-entrainment into the near-nozzle region increases the oxidation of the pyrolysis species of the fuel and reduces soot formation rate. Further, the additional dilution also reduces peak temperatures and consequently leads to lower concentrations

ferences

nnon S M, Queiroz M 1994 Time resolved temperature measurements in an elliptic cross-
section, turbulent diffusion jet flame. *ASME Winter Annual Meeting*, Chicago, IL, ASME vol.
HTD 296, pp 37–45
ghton D G 1973 Instability of an elliptic jet. *J. Fluid Mech.* 59: 665–672
llahalli S R 1977 Effects of diluents on the flame structure and radiation of propane jet flames
in a concentric stream. *Combustion Sci. Technol.* 9: 147–160
llahalli S R, Prabhu N 1990 Differences in the structure of lifted gas jet flames with laminar
base over circular and elliptic nozzles. *Emerging Energy Technology Symposium*, New Orleans,
LA, ASME Paper No. 90-Pet-14
llahalli S R, Savas O, Huang R, Rodriguez Azara J L 1987 Flow structure of a lifted gas
jet flame in the hysteresis region. *Twenty-first symposium on Combustion* (Philadelphia: The
Combustion Institute) pp 1463–1471
llahalli S R, Khanna T, Prabhu N 1992 Diffusion flames of gas jets issued from circular and
elliptic nozzles. *Combustion Sci. Technol.* 86: 267–288
tmark E, Schadow K C, Wilson K J 1985 The mean and turbulent structure of noncircular jets.
AIAA Conference, Colorado Springs, CO, AIAA Paper No. 85-0543
C M 1986 Mixing processes in free shear layers. *24th Aerospace Sciences Meeting*, Reno,
NV, AIAA Paper No 86-0234
C, Gutmark E 1987 Vortex interaction and mass entrainment in a small aspect ratio elliptic
jet. *J. Fluid Mech.* 179: 383–405
ssain F, Hussain H S 1989 Elliptic jets: part 1; Characteristics of unexcited and excited jets. *J.
Fluid Mech.* 208: 257–320
mal A 1995 *Turbulent diffusion gas jet flames from circular and elliptic nozzles*, Ph D disser-
tation, University of Oklahoma, Norman, OK
mal A, Gollahalli S R 1993 Effects of noncircular fuel nozzles on the pollutant emission charac-
teristics of natural gas burners for residential furnaces. *International Joint Power Conference*,
ASME vol. FACT 17, pp 41–50
anna T 1990 *Flow structure of jets and flames over elliptic nozzles*, M S thesis, University of
Oklahoma, Norman, OK
lluri P 1992 *The study of the interaction of noncircular venturi inlets with circular fuel jet in
the inshot burners of natural gas furnaces*. M S thesis, University of Oklahoma, Norman, OK
shigoe S, Gutmark E, Schadow K C, Tubis A 1988 Instability analysis on noncircular free jets.
26th Aerospace Sciences Meeting, Reno, NV, AIAA Paper No. 88-0037
panikolau N, Wierzba I 1995 Effect of burner geometry on the blowout limits of jet diffusion
flames in a co-flowing oxidizing stream. *Emerging Energy Technology Symposium*, Houston,
TX, ASME vol. PD-66, pp 29–36
panikolau N, Wierzba I 1996 Effect of burner geometry and fuel composition on the stability
of a jet diffusion flame. *Emerging Energy Technology Symposium*, Houston, TX, (in press)

Prasad A, Gundavelli S, Gollahalli S R 1991 Characteristics of diluent-caused lifted gas jet flames. *J. Propulsion Power* 7: 659–667

Quinn W R 1989 The turbulent free jet issuing from a sharp edged elliptic slot. *27th Aerospace Sciences Meeting*, Reno, NV, AIAA Paper No. 89-0664

Rao A V 1994 *Effects of venturi inlet geometry on combustion characteristics of inshot burners used in residential gas furnace systems*. M S thesis, University of Oklahoma, Norman, OK

Schadow K C, Wilson K J, Lee M J 1984 Enhancement of mixing inducted rockets with elliptic gas generator nozzles. *20th Joint Propulsion Conference*, Cincinnati, OH, AIAA Paper 84–1260

Scholfield D A, Garside J E 1949 Structure and stability of diffusion flames. *Third International Symposium on Combustion*, pp 102–109

Sforza P M, Steiger M H, Trentacoste N 1966 Studies on three-dimensional viscous jets. *AIAA J.* 4: 800–806

Subba S 1995 *Effects of burner exit geometry on combustion characteristics of inshot burners used in residential gas furnace systems*. M S thesis, University of Oklahoma, Norman, OK

Subba S, Gollahalli S R 1995 Flame structure and pollutant emission characteristics of noncircular partially premixed laminar gas jets. *Third Asian Pacific International Symposium on Combustion and Energy Utilization*, Hong Kong

Parallel power paths and compactness of gear transmissions

K LAKSHMINARAYANA†

Mechanical Engineering Department, Indian Institute of Technology, Madras 600 036, India

Abstract. Developments in mechanical power transmission have hinged around better materials and processes, new design ideas and effective new application of enduring scientific design principles and ideas. The current presentation aims at showing how deliberate measures to optimize load distribution have increased the compactness and capabilities of present day gear transmissions. Exploitation of parallel power paths, packing of gears in annular spaces and the effective use of epicyclic action by reducing the number of gear meshes and substituting rolling friction are explained.

Keywords. Mechanical power transmission; load distribution; parallel power paths; compact gear transmission.

1. Introduction

Transmissions serve primarily to convert and transmit motion and torque from a power source in a suitable manner for application. The source almost invariably provides uniform rotation. When it is converted into a non-uniform rotation, the device used is a non-uniform motion mechanism or a *non-uniform transmission*. Otherwise we have a *uniform transmission* or simply a *transmission*.

While there is a great variety of basically different transmission systems including hydrostatic transmissions and electric drives, we confine ourselves to purely mechanical systems here. In endeavouring to show how certain basic design principles play an important role in the modern development of compact transmissions, we confine ourselves mostly to gearing and indeed to only parallel-axis arrangements.

2. Basic design objectives

In transmitting power through a fairly high reduction, compactness is a basic requirement of steadily increasing importance. High efficiency then becomes essential even for small power units as the heat dissipation problem becomes more acute.

There are many other objectives such as reliability, cost effectiveness, ease of mainte-
nance, low noise levels, and ability to withstand critical environments. Keeping in view
the basic aim of this presentation as stated in the introduction, we confine ourselves to the
question of compactness and problems connected with the same.

3.   Load, stress and strength

Out of this triplet of basic factors forming the foundation for compact designs, *stress*
is decided mainly by form and size, while *strength* represents the role of material and
treatments. Great care is taken to minimize the stress and maximize the strength. A very
significant aspect of the *load* factor in realising the modern compact tranmissions is that of
load distribution/equalization. This is however often lost sight of. In achieving a satisfac-
tory level of load distribution, both a carefully developed application of design principles
and the availability of sophisticated manufacturing facilities have played a pivotal role.

   Design for effective load distribution thus becomes the main thrust of the current pre-
sentation. We limit ourselves to spur and helical gearing.

4.   Transverse distribution of load

This refers to the distribution of tooth load between two or more pairs of gear teeth
simultaneously in contact. Manufacturing accuracy (particularly in the form of low base
pitch error) is naturally the first requirement for a good distribution of load. A high contact
ratio (total contact ratio for helical gears) represents the theoretical occurrence of several
simultaneous tooth-pair contacts. A large number of teeth, a large tooth height and a low
pressure angle contribute to a high contact ratio. A large tooth height contributes to tip load
leverage (for tooth root stress) but the tip load itself is reduced due to the high contact ratio.
A more flexible tooth further helps in load distribution. A low pressure angle contributes to
the basic all-load pitch-point contact stress but the pitch-point load itself may be reduced
by the high contact ratio. The use of a large number of teeth on a given pitch cylinder
is mainly limited by the bending strength since the tooth size is reduced. The tooth form
however improves against the bending stress. Finally a large contact ratio generally reduces
the noise level significantly.

   Optimization of the number of teeth, pressure angle, tooth height and tooth profile
modifications represents one way of improving the compactness of the basic spur gear
pair. This qualitative discussion is to highlight the load distribution aspect.

5.   Longitudinal distribution of load

This refers to the distribution of load along the length of the gear tooth. Manufacturing
accuracy (particularly in the form of low tooth helix error) is naturally the first requirement
for a good distribution of load.

   In terms of practical requirements, compactness consists not only of low volume of space
occupied but also involves achieving a low radial dimension, i.e. a low centre distance. This
involves the permissibility of a relatively large face width. A larger face width however
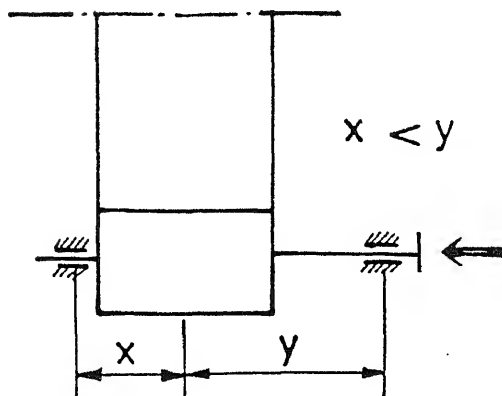
Figure 1. Straddle-mounted pinion situated near the bearing at the non-torque end of the shaft.

involves not only a possibly larger inaccuracy but also a larger bending and torsion of the pinion. Overhanging pinion is therefore avoided as far as possible and the pinion is straddle-mounted in bearings that are as close to the pinion as possible. The torsion effect on the pinion can be neutralised by the bending effect to some extent if the pinion is situated nearer the bearing at the non-torque end of the shaft. This is indicated in figure 1.

In the case of multistage gearing there is the possibility of a pinion on an intermediate shaft receiving power from both the ends. This is illustrated in figure 2. Maldistribution due to the torsion effect can be minimized in this fashion.

To improve the load distribution further, tooth helix modification has to be resorted to. This can be most advantageously used when the direction of major torque application is fixed and the magnitude of working load is fairly certain. The tooth helix correction may be made based on the most predominant load level or the average load. The correction for a symmetrically mounted pinion with one-sided torque application is illustrated in figure 3. In practice, an approximation by a helix angle correction or a single-sided crowning will often suffice. When both faces of a tooth are being corrected, the middle one-third of the tooth length may be left uncorrected (i.e. straight for spur gears) or linearly corrected. This is to help tooth helix inspection, as shown in figure 4a. When only one side is corrected (e.g., lifting equipment), circular arc form crowning can be ground on that side as shown in figure 4b. The other side is left uncorrected for tooth helix inspection. The unsymmetrical
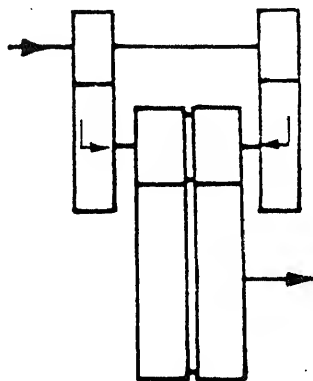
Figure 2. Pinion on an intermediate shaft receiving power from both the ends in multistage gearing.
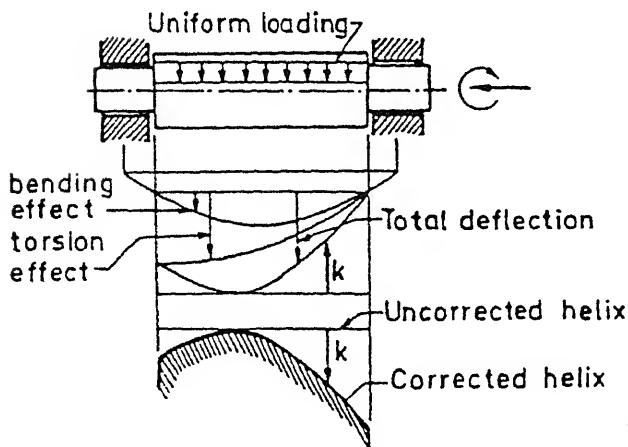
Figure 3. Illustration of correction for a symmetrically mounted pinion with one-sided torque application.

crowning in figure 4b is the result of the approximate linear correction already referred to earlier and symmetrical crowning is introduced to counter manufacturing inaccuracy. In case of doubt it is better to go in for simple tooth-end relief since over-correction can prove to be detrimental.

6. Load distribution between different gear pairs: Parallel power paths

Figure 5 shows dual trandem gearing, also called the locked train. There are now two parallel power paths as shown in figure 6. Maldistribution of load between the two power paths can result from tooth errors (especially pitch errors), errors in the relative angular position of gears ($3'$ w.r.t. $2'$ as against 3 w.r.t. 2 in figure 5), errors in the circumferential location of the intermediate gears as planets around the sun gears (1 and 4 in figure 5) or errors in δ_1 and δ_4 in figure 5.

Figure 7 shows a simple planetary gear drive with three simple planets. The three parallel power paths are also shown in the same figure. Compactness is achieved by packing the annular space between the sun gear 1 and internal gear 3 with several planets. Up to 6 planets are used, obtainable speed ratio being reduced as the number of planets is increased. The
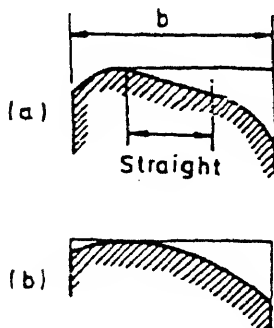
Figure 4. (a) Middle one-third of tooth length is left uncorrected for tooth helix inspection when both tooth faces are being corrected. (b) Only one side is corrected and circular arc form crowning is ground on that side.
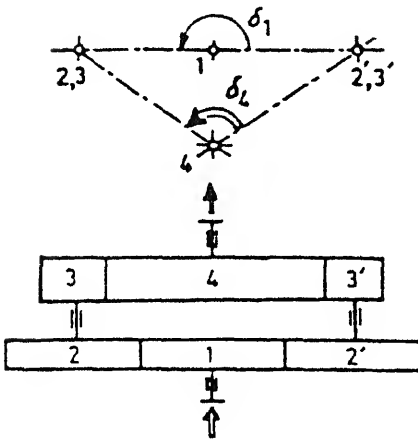
Figure 5. Dual tandem gearing, also called locked train.

ompactness is however conditional on the achievement of an equitable distribution of oad between the parallel power paths.

There are basically three different ways of proceeding towards load equalisation amongst arallel power paths:

(i) manufacturing accuracy,

(ii) flexibility, and

iii) kinematically correct design (statical determinacy).

Some flexibility is always present, due to the elasticity of the machine elements. If there vere no flexibility, only one power path would carry the entire load at any particular mo- nent, depending on the relative magnitudes of the tooth spacing errors involved. Flexibility nables the gaps in the other power paths to close when the total load reaches a certain evel. The additional load is carried equally by all the power paths.

Thus, with the same flexibility and ultimate load intensity, the lower the spacing errors, ne more equitable is the distribution of load. Given the same spacing errors and the same ltimate load intensity, the higher the flexibility the more equitable is the load distribution mongst the parallel power paths. High flexibility, however, can cause vibration problems, eyond a point. It is therefore essential that a high degree of manufacturing accuracy be naintained, consistent with cost effectiveness. It may then be found, especially for small nits, that the flexibility already present in the system is enough.

With a kinematically correct design, there is statical determinacy in the system. That s, the loads in the individual power paths are known (independent of the spacing errors vhich vary through a cycle and from unit to unit). The loads may not always be exactly qualized (this is further influenced by friction which is somewhat indeterminate).

Figure 6. Illustration of two parallel
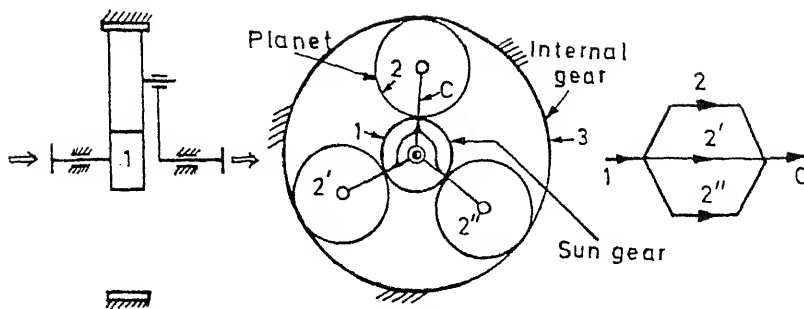
Figure 7. Schematic of a simple planetary gear drive with three planets. The three parallel power paths are also shown.

The standard kinematic solution to the load distribution problem for the three-planet drive of figure 7 is to 'float' the sun gear 1. This is normally achieved by a gear coupling between the motor and the sun gear as shown symbolically in figure 8. The coupling transmits a pure torque without imposing a lateral force on the sun gear 1. Ignoring friction, the three total tooth normal forces between the sun gear and the three planets are equalized when the planets are spaced at 120° on the carrier. This load equalization is greatly dependent on effective lubrication of the gear coupling. This lubrication can easily fail since the internal movements in the coupling engendered by the varying tooth spacing errors in the gearing proper are rather small.

The sun gear centre moves in small loops, occasioned by the spacing errors. These movements can generate very significant inertia forces at high speeds with their dynamic effects on the meshes with the three planets. It is thus seen that a high degree of manufacturing accuracy is still essential.

Instead of the sun gear, the carrier may be 'floated' by providing a gear coupling at the output. The torque at the carrier being higher, the gear coupling will be larger.

Similarly, the internal gear may be floated. Coupling lubrication is then more involved since there is no rotation to help. Some proven high-power high-speed drives float both the sun and the internal gear. The inertia effects mentioned earlier are thus mitigated.

Kinematically correct designs for more than three planets have also been developed. Figure 9 shows a design for a 5-planet drive schematically. Here the carrier C is floated by means of a gear coupling at the output (coupling not shown). The sun gear is not floated.
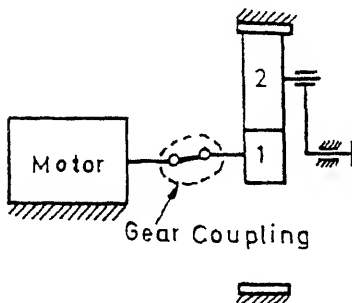
Figure 8. Symbolic representation of gear coupling between the motor and the sun gear.
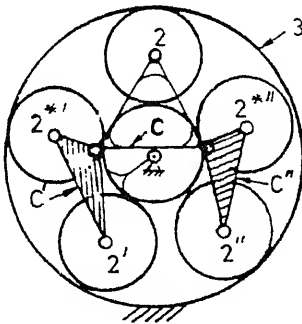
Figure 9. Schematic design for a 5-planet drive.

Auxiliary carriers C′ and C″ carry two planets each. The dimensions of the carriers are adjusted to provide as equal a distribution of load as possible amongst the planets and ensure static balance. The planets are equally spaced.

The above and similar developments are based on the planar model. While theoretical studies have been made with the three-dimensional model, a practical approach to the third dimension is to use crowning on some of the gears.

Kinematically correct designs for the dual tandem gearing of figure 5 are now considered. Out of a large number of design possibilities, some are indicated below.

(i) Float the driving pinion (gear 1) by gear coupling.

(ii) Mount the driving pinion (gear 1) on a long lever.

(iii) Guide the bearing housing, in which the intermediate shafts are mounted linearly, perpendicular to centre distance.

(iv) Use helical teeth for gears 1, 2, 2′. Make the two intermediate shafts 2–3 and 2′–3′ axially free. Balance the axial forces on these two shafts by means of springs or by means of the fluid connection as shown in figure 10.

(v) Introduce additional pinion 1′ with opposite helix (vis-a-vis 1) to mesh with 2′ and make the combined driving pinion 1–1′ axially free.

One of the guiding principles is to float only members with low inertia.

Some of the flexibility solutions to the load distribution problem are now indicated. In the dual tandem gearing of figure 5, torsionally flexible couplings may be interposed between gears 2 and 3 and between gears 2′ and 3′. Alternatively, the torsional flexibility

Half gear Coupling

3'

4

2'

Half gear Coupling

Torque Shaft **1**

2
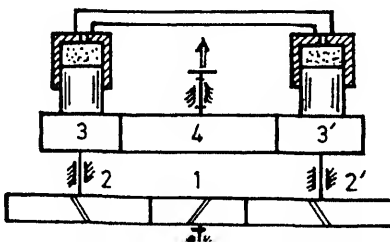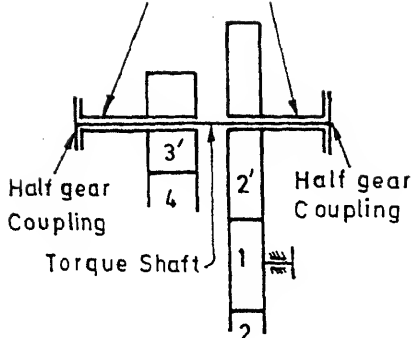
Figure 11. Increase of torsional flexibility of two intermediate shafts by means of the quill shaft design.

of the two intermediate shafts may be increased by means of the quill shaft design of figure 11. In the 3-planet drive of figure 7, the internal gear may be mounted on flexible material (or equivalent designs in metal).

7. Assembly conditions for parallel power paths

Closely linked with the statical indeterminancy of the parallel power path arrangements is the problem of assembly of gears of the additional power paths beyond the first. In figure 5, gear meshes 1–2, 3–4 and 1–2' can first be assembled without any problem. Gear 3' can then be assembled with gear 4 only at particular angular positions of 3' with respect to 2'. If one of these relative orientations is used, then the gears can be assembled. It is normally desired however to make units 2–3 and 2'–3' of identical specifications. The angles δ_1 and δ_4 may then be suitably chosen or the number of teeth adjusted. In the planetary gearing of figure 7, the sum of the number of teeth on the sun gear and on the internal gear should be divisible by the number of planets, if the planets are to be equally spaced.

8. Compactness through epicyclic action

When a fairly high ratio of speed reduction is required, use of single-stage gearing will make the wheel very large. Two, three or more stages are thus introduced. The space requirement however can still be too large. Two or three simple planetary stages may be used and the coaxial input and output exploited. The compactness is otherwise due to the multiple planets in the individual stages.

In a simple epicyclic gear train consisting of two coaxial gear shafts a,b and a carrier shaft c, we can choose the tooth numbers such that the carrier-reference speed ratio i_{ab} is very near to $+1$, b being fixed. We can then obtain an epicyclic speed reduction ratio of $I_{ca} = 1/(1 - i_{ab})$. If $i_{ab} = 1.02$, the speed reduction ratio from carrier c to co-axial gear a will be $= 1/(1 - 1.02) = -50$. Since the ratio i_{ab} is very nearly 1, the tooth numbers are small and the drive is very compact for the large ratio demanded. The efficiency however goes down severely. With a tooth power loss factor of 0.005, the efficiency will be only 66% and becomes just 50% if the tooth power loss factor is 0.01. With a low efficiency, large
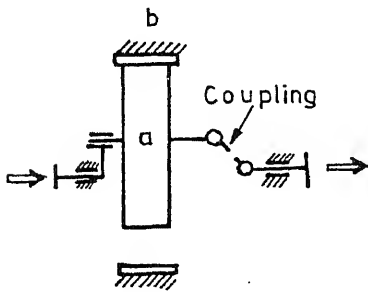
Figure 12. The 'open' train – reduction in the number of gear meshings from 2 to 1 in a power path.

heat is generated and the particularly compact unit is not equipped to dissipate the same. To improve the efficiency, one step is to reduce the number of gear meshes in a power path from 2 to 1. The result is the 'open' train shown in figure 12. Here the absolute rotation of the planet is the output motion. This is 'brought down' to the axis of the co-axial unit by means of a constant velocity coupling using rollers. To obtain a large reduction, the number of teeth on the planet and internal gear differ by a small number, creating the problem of tip-to-tip interference. Special involute tooth proportions are needed. Using rollers in place of involute teeth on gear b and a tooth difference of 1, the tip-to-tip interference problem is circumvented by the rollers never leaving the mating planet. The demand for a low tooth power loss factor is simultaneously satisfied. (Due to rolling friction as well as low roller velocities; the latter are due to the internal–external mesh of almost +1 ratio.) This is the *cyclo-drive* system. Instead of the rollers being rotated about axes fixed on the ring gear (internal gear), they may be housed in a separate cage and interposed between concave pockets on the ring and the planet profile. This is the *quadrant drive* system.

Vibration – A tool for machine diagnostics and condition monitoring

K N GUPTA

Department of Mechanical Engineering, Indian Institute of Technology, New Delhi 110 016, India
Present address: Institute of Engineering & Technology, Rohilkhand University, Bareilly 243 001, India

Abstract. Vibration is an effective tool in detecting and diagnosing some of the incipient failures of machines and equipment. The present paper deals with the basic principles, which may help in identifying its diagnostic ability, the scope of its diagnostic capabilities, the instrumentation in vogue for its monitoring and the state-of-the-art of the monitoring techniques and programs. A few case studies are also given to illustrate how machine troubles/failures are diagnosed with the help of vibration signatures.

Keywords. Vibration analysis; vibration signatures; machine diagnostics; condition monitoring.

1. Introduction

The subject of vibration generally deals with methods to determine the vibration characteristics of a system, its vibratory response to a given excitation and the means to reduce the vibration. Reference is however hardly made to its diagnostic ability. A vibration signature measured on the external surface of a machine or a structure contains a good amount of information, which, if properly interpreted, can reveal the running condition of the machine. It may be regarded as one of the languages through which a machine tells us of its ailments. This information is useful in making suitable decisions regarding machine or plant maintenance. Attempts have been made for the last two decades to exploit this feature of vibration signature for the benefit of machine condition monitoring, and the result is a well proven technology now in vogue. This success has been possible due to the tremendous developments made by specialists in the areas of instrumentation, electronics and computers. Books by Collacott (1973, 1977) are often referred to, whenever a topic on condition monitoring and fault diagnosis is discussed. Articles by Gupta (1986, 1990)

out of maintenance, only when the condition of the equipment requires it as a safeguard from incipient failure. This strategy helps in reducing maintenance costs and chances of unscheduled breakdown or shutdown of the equipment. An improvement in productivity is inevitable. However, this requires sophisticated electronic instruments to monitor the condition of the equipment. Hence a knowledge of such instruments is essential.

Before a brief description of such instruments is given, it is thought appropriate to list the diagnostic capabilities of a vibration signature vis-a-vis machine ailments and to enunciate the basic principles involved, which may help in exploring further such capabilities. The present paper provides a glimpse of the state-of-the-art of this diagnostic technique and the possible future trends in its application to plant and machine maintenance. Some case studies are also included to illustrate how a fault could be diagnosed through vibration signatures.

2. Basic principles involved

The technique of condition monitoring through vibration is built up on the following principles

(1) Any malfunctioning or deterioration in the operation of a machine component gives rise to increase in vibration level.

(2) Vibration emanating from a component consists of certain frequencies depending upon its nature of operation. This frequency information does not get changed or lost during transmission of vibration, however, their vibration level may be attenuated.

(3) Mixing of different vibrations does not cause any loss of the individual's frequency information.

(4) Every individual component or system has its own frequency, called its natural frequency, which changes only when the system parameters get affected.

Figure 1 illustrates nicely some of the principles enunciated above. The unbalance in the coupling generates vibrations with frequency equal to the rotational speed, the bearing generates vibration of frequency depending on the number of balls and the gear
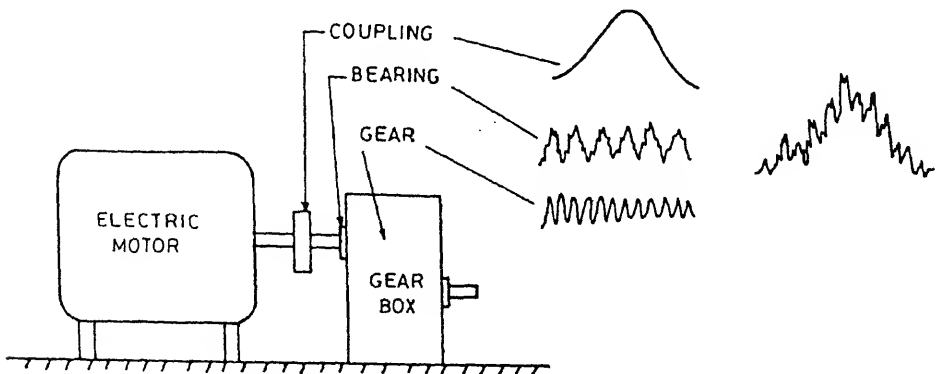
Figure 1.   Vibration generation and transmission.

nerates vibrations of tooth-meshing frequency. These vibrations mix with each other
ring transmission, and the mixed signal is picked up by the transducer. The mixed
nature has all the frequencies of the individual vibrations, i.e. no information with
ard to frequency is lost in mixing or transmission.

Diagnostic capabilities of a vibration signature

vibration signature taken from an appropriate location in a machine can reveal the pres-
ce of the following machine defects: Imbalance, misalignment, imperfect foundation,
chanical looseness, rubs, antifriction bearing defects, faults in belt drive, faults in gears,
eve bearing looseness, oil whirl, blades/vanes defects, local resonances, etc.

In general, any machine defect, which alters the dynamic behaviour of the machine can
detected through a vibration signature.

Instrumentation and diagnostic parameters

r carrying out vibration monitoring, suitable instrumentation is required and, depending
the nature of the diagnosis, appropriate processing parameters, also called diagnos-
parameters, should be identified. This section deals with these aspects of vibration
onitoring.

Selection of proper instrumentation for vibration monitoring is an important feature.
vibration measuring system consists of a sensor or transducer, a signal processing unit
d a display unit or recorder. The transducer is used to convert mechanical vibration
o an electrical signal. The signal processing unit conditions, amplifies, integrates or
fferentiates and analyses the signal for its frequency spectrum. Finally, the display or
corder unit presents the vibration characteristics in an interpretable form.

The three vibration transducers generally used have the following practical measurement
plitude ranges:

- the proximity displacement transducer $100 : 1$
- the velocity pickup $1000 : 1$
- the piezoelectric accelerometer $30 \times 10^6 : 1$

us, from this angle, the piezoelectric accelerometer should be the best choice. Further,
is lighter and has better frequency range for application. These days accelerometers
used as sensors in most vibration monitoring systems. Some instrument manufacturing
ms, particularly Brüel & Kjaer of Denmark, have developed a well-perfected technology
vibration monitoring with these sensors.

Proximity displacement transducers are generally used in case of rotating machines hav-
g their rotors mounted on oil bearings. Displacement probes, generally two in number at
ch bearing support, monitor the orbital motion of the rotor. It is the shape and nature of
ese orbital paths which provide the clues for impending defects in the rotor. Defects like
balance, misalignment, oil/shaft whirl and rotor cracks are easily detected. M/s Bently,
vada, USA are the pioneers in developing a technology with eddy current probes as trans-
cers. The turbo-generator sets of many powerplants are equipped with these instruments.

Table 1. Time-based vibration analysis (Tranter 1989).

Overall level (r.m.s.) measurements	Most common vibration measurement in use. Most simple and inexpensive type of measurement. Greatest limitation is the lack of sensitivity and information available in the data. Unless a problem is severe, r.m.s. may not change significantly.
Peak level detection	Particularly useful for monitoring the change in the amount of impulsiveness, possibly due to increased bearing damage. This method is not 100% reliable, as other effects can also increase the peak level of a signal.
Crest factor	The crest factor (sometimes called the impact index) is the ratio of the peak level to the r.m.s. level. This method also has limitations.
Shock pulse & spike energy	Basically a measure of the vibration level at the bearing resonance, usually above 30 kHz. Widely used, however, concern has been expressed as the reading can decrease in later stages due to a reduction in impulsiveness, and other conditions, such as turbulence and cavitation in pumps, can give false readings.
Kurtosis	Statistical parameter, derived from the statistical moments of the probability density function of the vibration signal. The Kurtosis technique has the major advantage that the calculated value is independent of load or speed variations.
Demodulation (envelope detection)	Often the bearing signals are swamped by more dominant low frequency signals. This method, which can be implemented as a Hilbert transform, filters out low frequency signals, leaving a clean signal dominated by the bearing frequencies.
Phase	Phase indicates the relative timing between two points. It is used in balancing and is useful when diagnosing imbalance, misalignment, looseness and other cases.
Time waveform	Using an oscilloscope, it is possible to view the waveform of the vibration. Difficult to use in isolation, it can be very helpful tool in combination with others.
Orbits	Taken using two channel oscilloscope connected to proximity probes. More recently, they have been derived from a pair of frequency spectra. The major benefit is that they show the relative motion of the dominant vibration of the shaft.

Choice of accelerometer is not restricted to measurement in the acceleration mode only, because an acceleration signal can be converted into velocity or displacement with no loss of accuracy. The best mode to be used for measurement is the one that gives a spectrum with little or no slope, as this takes up the least dynamic range in a measuring system. It is found that velocity is the best mode.

A vibration signature, as obtained by the accelerometer or any other transducer, is a time-base signal which may be processed for its overall level, peak level, phase, spike energy etc. Some of the diagnostic indicators can be evaluated by making time-based analyses. Table 1 summarizes these indicators, which can be adopted as useful parameters in trend analysis.

A vibration signature may be a combination of various harmonics, which is obtained by processing the signal through a frequency analyzer. Each part of the frequency is a potential

Table 2. Common faults and frequencies (Tranter 1989).

Frequency	Possible cause	Comments
1× rpm	Imbalance	Steady phase that follows transducer. Can be caused by load variation, material build-up or pump cavitation
	Misalignment or bent shaft	High axial levels, 180° axial phase relation at the shaft ends. Usually characterized by high 2× rpm
	Strain	Caused by casing or foundation distortion, or from attached structures (e.g. piping)
	Looseness	Directional changes with transducer location. Usually high harmonic content and random phase.
	Resonance	Drops off sharply with change in speed. From attached structures
	Electrical	Broken rotor bar in induction motor. 2× slip frequency sidebands often produced
2× rpm	Misalignment or bent shaft	High levels of axial vibration
Harmonics	Looseness	Impulsive or truncated waveform; large number of harmonics
	Rubs	Shaft contacting machine housing
Sub-rpm	Oil whirl	Typically 0.43–0.48 of rpm; unstable phase
	Bearing cage	Fundamental train = 0.5× rps [1-(ball dia)/(pitch dial) × COS (contact angle)]
$N \times$ rpm	Rolling element bearings	Inner race = 0.5 × # balls × rps [1+ (ball dia)/(Pitch dia) × COS (contact angle)]
		Outer race = 0.5 × # balls × rps [1-(ball dia)/(pitch dia) × COS (contact angle)]
		Ball defect = 0.5× (pitch dia)/(ball dia) × rps [1-(ball dia)/(pitch dia) × COS (contact angle)^2]
		Usually modulated by running speed
	Gears	Gearmesh (# teeth × rpm); usually modulated by running speed
	Belts	Belt × running speed and 2× running speed.
	Blades/vanes	# Blades/vanes × rpm; usually present in normal machine. Harmonics usually indicate that a problem exists
$N \times$ powerline	Electrical	Shorted stator; broken or eccentric rotor
Resonance		Several sources, including shaft, casing, foundation and attached structures. Frequency is proportional to stiffness and inversely proportional to mass. Run-up tests and modal analysis are useful in this area

indicator of the machine condition. Spectrum analysis is the most powerful technique for diagnostic study. The underlying principle is that each operating component of the machine generates identifiable frequencies. Thus, changes in the vibration level at a given frequency can be related directly to the concerned machine components. Based on the basic knowledge associated with the nature of machine operation and the principles enunciated in § 2, one should be able to calculate the frequencies consequent to the impending faults. In the literature, charts supplying this information are available. Table 2 is one such chart giving a summary of fault frequencies and the relevant faults (Tranter 1989).

Sometimes, a change in the individual frequency component does not bring about any
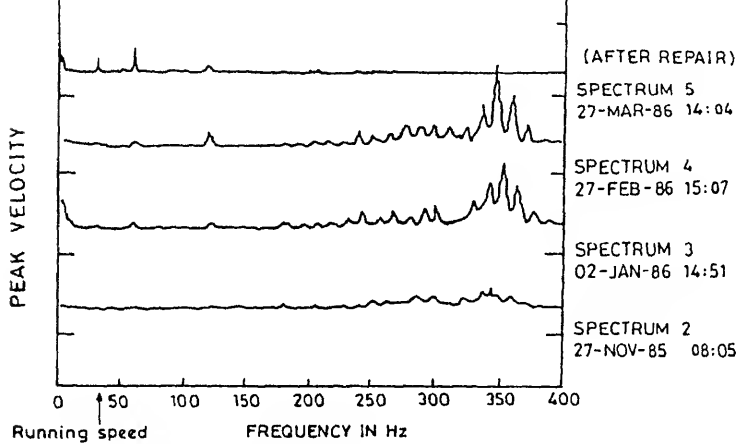
Figure 2. High pressure water pump bearing failure.

that the bearing condition was steadily deteriorating. In this particular case, balls and races were found badly spalled and the cage was fractured at two places.

There are a number of indices derived from the frequency spectrum. They are very useful parameters for trend analysis. Table 3 lists these parameters and describes their diagnostic characteristics.

For trend analysis to be performed, a reference base signature of the machine in good condition is needed. To ascertain whether·the machine is in good condition, three or four machine signatures are taken at weekly intervals under the same operating condition. If no change is noted in the vibration spectrum, the machine's internal condition must be reasonably stable and the machine can be considered in good health. The first spectrum can be used as a base line spectrum for trend analysis.

Trend monitoring strategy is adopted where it becomes difficult to generate acceptable vibration levels. The underlying idea in trend monitoring is that the rate of deterioration in vibration level of the machine increases rapidly near breakdown. The vibration signature is recorded regularly, and the values for the chosen indices as listed in tables 1 and 3 are determined, and the trend noted. The lead time before the degradation reaches breakdown is one of the main advantages of the trend monitoring. During the lead period maintenance actions are planned.

5. Vibration monitoring programmes

The monitoring programmes can be classified into three levels of sophistication. This is reflected in the speed with which they can detect the faults and provide information for locating them.

The simplest system consists of an accelerometer and a vibration meter (figure 3a). It measures the vibration level over a specific wide-frequency range. Measurements are compared with the established reference values for each machine. If such reference values are not generated for each machine, they can be taken from the general standards like

ble 3. Frequency based vibration analysis (Tranter 1989).

ectrum	Derived from the vibration waveform by performing a 'fast Fourier transform' (FFT). Given that the running speed of the machine is directly proportional to the frequency measured, it is possible to relate peaks in the spectrum to the machine components
aterfall plot	A waterfall plot (also known as spectral map and cascade plot) is a three-dimensional representation of spectra, usually with time as the third dimension
epstrum	The FFT of the logarithm of the power spectrum. Used to high-light periodicities in the spectrum. Useful in bearing and gear-box analysis
fference spectra	By mathematically subtracting two spectra, changes in level are easily identified. Fault frequency analysis is performed to relate the frequencies to the machine components. Does not cope well with running speed changes
oot mean square of difference	The rms of the difference between the current spectrum and the baseline, and the current spectrum and the previous spectrum, have both been found to be useful trending parameters
atched filter	Another method of trending the difference between vibration spectra. Found to be a reliable trending parameter. Quantifies the differences between spectra by summing the squares of the corresponding amplitude ratios in the spectra and taking the logarithm of the result

DI 2056. This is the cheapest system and the operator does not require much training. achine condition is evaluated in the field from minimum data. The system however has me limitation. If the change in vibration level is from more than one source, the method nnot diagnose the actual source of trouble. For diagnosing the source of trouble some struments are required to perform the time-based analysis to obtain the parameters, as entioned in table 1. Some of these facilities are provided in the vibration meter. In the ecific case of rolling element bearings, and where vibrations from other sources do not ominate, early warning of bearing deterioration can be obtained. The meter is switched to easure simultaneously both the peak and rms vibration levels up to 10 kHz. At the earlier age of fault development in the bearings, very little change in the rms level is noted, though the peak level increases significantly. By evaluating the crest factor (peak/rms tio) many rolling element bearing faults can be detected at an earlier stage.

Fault detection at an earlier stage together with diagnosis is also possible, when using system which can perform frequency analysis. Full frequency analysis and spectrum ot-out are done on the spot for each monitoring point. The system is shown in figure 3b. urrent spectra are compared manually with the recorded reference spectra to reveal tell-le increases in the level of individual frequency components. Reference (baseline) spectra e recorded for each monitoring point and transferred to transparent sheets. Subsequent ectra are placed under the reference sheets, where any difference will be immediately pparent. As the levels of certain frequency components begin to grow, they are plotted a level versus time chart, so that failure trends can be predicted. This enables the fault be diagnosed, spare parts to be ordered and repair to be scheduled conveniently. Some
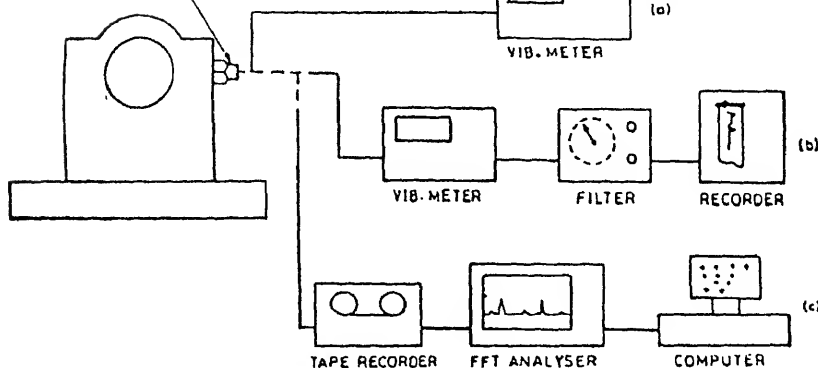
Figure 3. Vibration monitoring systems. (a) Simple system for vibration measurement; (b) system for frequency analysis; (c) system for computer-aided off-line trend monitoring.

users perform simple wide band monitoring on a regular basis and employ the frequency analyzer when sufficient changes in the vibration level are noted. While analysis at this stage will aid diagnosis of the developing fault, the early warning and trend benefits of regular analysis are not available in this case.

As the number of monitoring points increases, a computer-aided spectrum comparison system will be the most economical solution. Vibration signatures from each machine are collected on an instrumentation tape recorder. They are analyzed on a real-time frequency analyzer in the office, and the current spectra are compared with the baseline spectra under semi-automatic control from a desk-top computer coupled to the analyzer (figure 3c). Advanced programmes aid fault diagnosis and trend monitoring. Narrow-band frequency analysis on a linear frequency scale as provided by the FFT analyzer gives excellent display of harmonic and side-band frequency components, a particularly valuable information for diagnostic purposes. Where the analyzer includes a zoom facility, any part of the spectrum can be expanded to further enhance the details of individual components.

6. Permanent monitoring system

Such a vibration monitoring system is permanently installed on a specific machine and continuously watches its condition. Its function is to give immediate warning of any sudden changes in the condition of expensive non-duplicated machinery, whose continuous operation is vital to the production process. Faulty conditions are detected immediately or within minutes of occurrence and trigger alert or alarm signals in the plant control room, so that appropriate measures can be taken before a catastrophic failure occurs. These systems are widely used in the power generation and petrochemical industries on turbines, feed pumps, gas compressors etc. A typical system is shown in figure 4.

In a basic system, a single module may continuously monitor vibration over a single specified frequency range. If preset limits are breached (e.g. minimum, alert and alarm), the system can trigger visual or audible alarms. Alternative options can provide for up to three individual frequency bands to be monitored simultaneously. Coverage of many
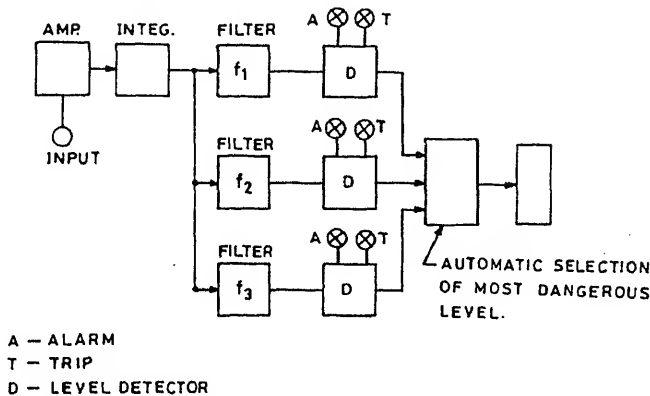
Figure 4.   Permanent monitoring system.

monitoring points can be provided by a multiplexer connected to the input of a single monitoring module. A multiplexer continuously steps through the chosen channels, dwelling at each channel for a preselected period before automatically moving on to the next. Several multiplexers can be connected to obtain the economic multichannel systems. Spectrum analyzers can be coupled to the monitoring system when fault diagnosis is required and, in the most sophisticated systems, computer-controlled fault detection is applied automatically to all monitoring points.

Main requirements for all permanent monitoring systems are extremely high operational reliability, long-term stability and immunity to adverse environmental conditions and irregularities which can cause false alarms. To safeguard from false alarm, these systems are sometimes provided with an automatic test system, so that the plant operator can immediately check whether the instrumentation is functioning correctly in the event of an alarm.

7.   Future trends in vibration monitoring and fault diagnosis

Future trends, which are visible in the present day practice, are given below.

7.1   *Permanently installed monitoring devices*

Machines and equipment with built-in monitoring devices are increasing day by day. Generating sets, steam and gas turbines, special purpose machine tools and production machines, refinery equipment are just some of the examples. It is foreseen that through increased and improved vendor relations in the near future, critical equipment will have installed monitors at the insistence of the customer or through the intention of the vendor. This trend is surely going to result in better reliability, availability and maintainability of the machinery.

7.2   Computers for data processing and management

manually. Further, it is true that the use of a number of parameters and techniques in combination will give the best indication of machine condition. This means that large amounts of data must be collected, analyzed and interpreted. Computers can help in a big way to manage such huge amounts of data and store useful information after data reduction. This information can be retrieved in no time for trend analysis. Use of computers is gradually increasing. In some cases, monitored data are manually fed to the computer, while in others it is automatically done with the use of multiplexers and A/D interfacing. The computer gives out alert reports, trend graphs and diagnostic information.

7.3   Expert systems

Fault tracing on the basis of monitored and analysed data is not always straightforward. It needs guidance from the experts who, based on their knowledge and experience, are able to pin-point the source of trouble. An expert, besides possessing knowledge of the subject, also has the ability to think in a progressively logical manner, which is a positive advantage in the world of problem-solving associated with plant and machinery faults. Perhaps this is where the area of 'artificial intelligence' has something to offer.

An expert system consists of a knowledge-base, an inference engine and a user interface. The knowledge-base contains a set of rules and facts that an expert system uses to solve a particular problem. The rules describe the relationship between the machinery problems and the corresponding symptoms. The inference engine contains control mechanisms to control the operation of the system and to infer information from the knowledge-base by using the rules to reason about the facts. The user interface enables communication between the user and the computer.

8.   Case studies

Three case studies are presented to show how the vibration signature could help in diagnosing the source of troubles.

8.1   Dryer foundation vibrations

In this case study (Gupta et al 1986), the problem was of excessive vibrations at the foundation block near the discharge end, and the objective was to determine the cause of these vibrations.

The dryer consists of a shell of 20 m length and 2.5 m diameter. It is supported on two foundation blocks, 12 m apart, through two pairs of rollers. A schematic view of the shell on the foundation block nearer to the discharge end is shown in figure 5. The dryer is meant for drying polypropylene (PP) granules. It is driven at a speed of 3.5 rpm by an electric motor through reduction gears and a pinion. The complete drive unit is mounted on the foundation block nearer to the intake end of the dryer. The PP granules passing through the dryer get dried by the steam passing through a number of tubes arranged along the shell length and near its circumference. A number of hammers are arranged along the
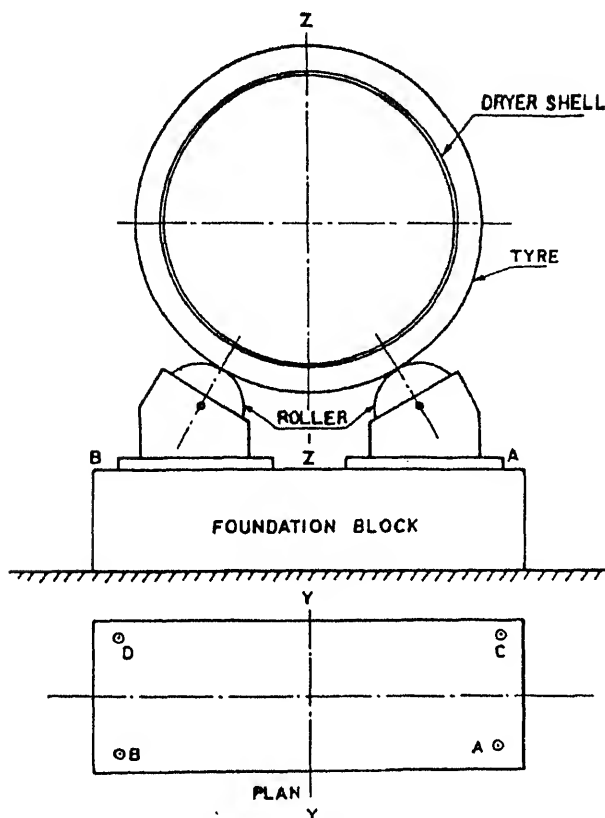
Figure 5. Schematic of the sectional view of the dryer on foundation and location of measuring points.

ircumference of the shell and near the intake end. These hammers fall on the shell surface nder gravity during shell rotation and thus tapping the shell surface, so that PP material hould not stick to the surface and choke the passage.

Vibration measurements were conducted on the foundation block. A time-base signal s recorded is shown in figure 6, and the corresponding frequency spectrum is shown in gure 7. The time-base record corresponded to a narrow-band signal showing the dominance of 6 to 7 Hz vibration. Of course, the amplitudes varied in random fashion. The requency spectrum also confirmed the presence of 5.2, 6.0 and 6.5 Hz harmonics. A simple heoretical analysis revealed that these frequencies were nothing but the natural frequencies of the foundation block in three modes. Further, it was noticed in the time-base record hat there was repetition of the similar pattern of the signal after every 4.0 s, corresponding o a frequency of 0.25 Hz, which happened to be the rotational speed of the roller. Zones f these repetition are marked in figure 6. Hence, it may be inferred that some irregularity n the roller surface caused this repeated phenomenon.

From the foregoing discussion it was concluded that the foundation block was getting xcited in its natural modes. This was possible if it was excited by a wide-band process ncompassing the frequencies of the natural modes. It was realized that the wide-band
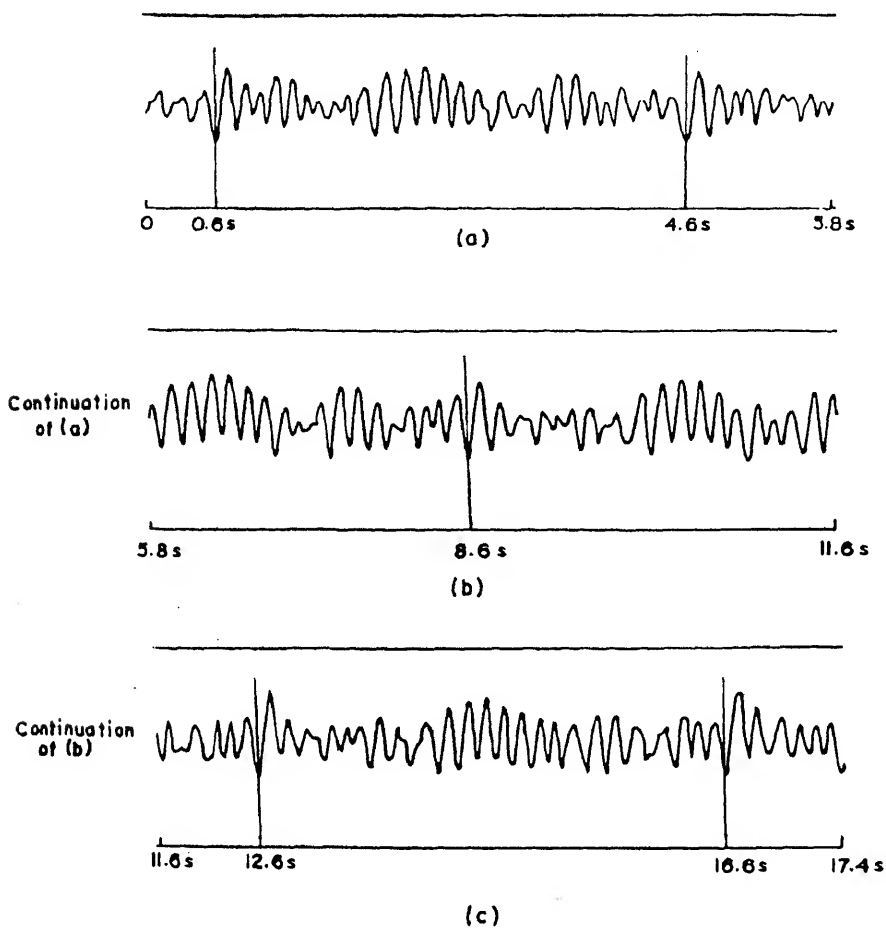
Figure 6.   Time-based record of the vibration at the foundation.
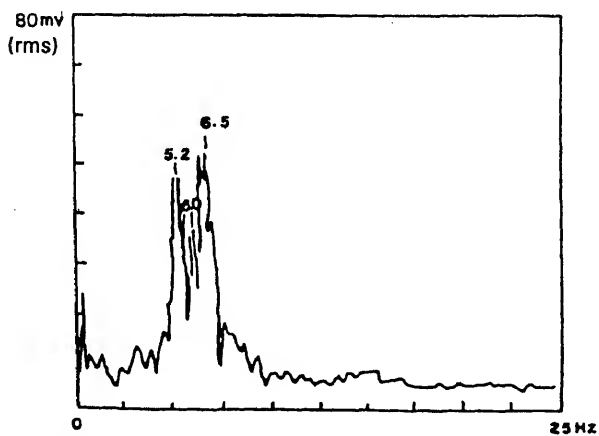
Figure 7.   Frequency spectrum of the foundation vibration.

These pits might have been formed in course of time due to excessive stresses and lack of lubrication. The pits were of different sizes and spread randomly over the surfaces. During operation of the dryer, if pits of minute sizes fell on the track of contact, the foundation had low level vibration, while in case of big pits, there was excessive vibration.

The following recommendation was made.

> "Pit formations are because of excessive contact stresses and lack of lubrication. Contact stresses can be reduced by increasing the area of contact. Sufficient lubrication can be provided by dipping the rollers in the oil bath, so that their surfaces are well smeared with oil before coming in contact with the tyre. If pit formations still persist, they should be repaired during the shut down period by pouring welding material and then grinding the spots."

The company did adopt the above measures and did not face any more problem of excessive vibrations at the foundation block.

8.2 *Excessive vibrations of the piping connecting booster pump (BP) and boiler feed pump (BFP)* (Gupta *et al* 1988)

While commissioning a super thermal power unit (500 MW) at a certain power station, vibrations of very large amplitude were observed in the piping connecting the BP and the BFP under certain conditions of operation, viz. safety valve blow-out operation. The region of these severe vibrations is marked in the characteristic curve of the BFP (figure 8). We were required to detect the cause of the trouble and also suggest a possible solution to the problem.

Figure 9 shows the relative positions of BP, BFP and the piping layout. Both BP and BFP are installed on the ground floor of the Power House, while the deaerator tank (DT) is situated at a height of about 26.5 m, thus providing a positive suction head of 30.5 m of liquid column to the BP. The water from DT goes to the BP through a barrel-type strainer having a mesh size of 0.2 mm. The BP is driven by an electric motor at a constant speed of 1493 rpm. The discharge from the BP is connected to the BFP through another barrel-type strainer with a mesh size of 0.5 mm. The BFP is driven by the same electric motor through gears and hydraulic coupling. Its speed can be controlled from 1490 to 5905 rpm by changing the position of the scoop in the hydraulic coupling. The discharge pipe of the BFP is mounted with a recirculation valve, which directs the excess flow to the recirculation piping leading to the DT.

We decided to record the vibrations of the piping at suitable locations under conditions of severe and non-severe vibration and to analyse these records in the laboratory for the frequency contents. We thought that a comparison of their vibration spectra would provide a clue to the identification of the source of trouble.

Figure 9 shows the four locations at which vibration signatures were recorded in the three directions. Signatures were also recorded at the BP suction pressure tapping (location 5 in figure 9), in order to have an idea of the frequencies of pressure pulsations in the water. Vibration measurements were carried out on two occasions. On both occasions, so-called severe vibration conditions could not be achieved. Vibration signatures were analysed for
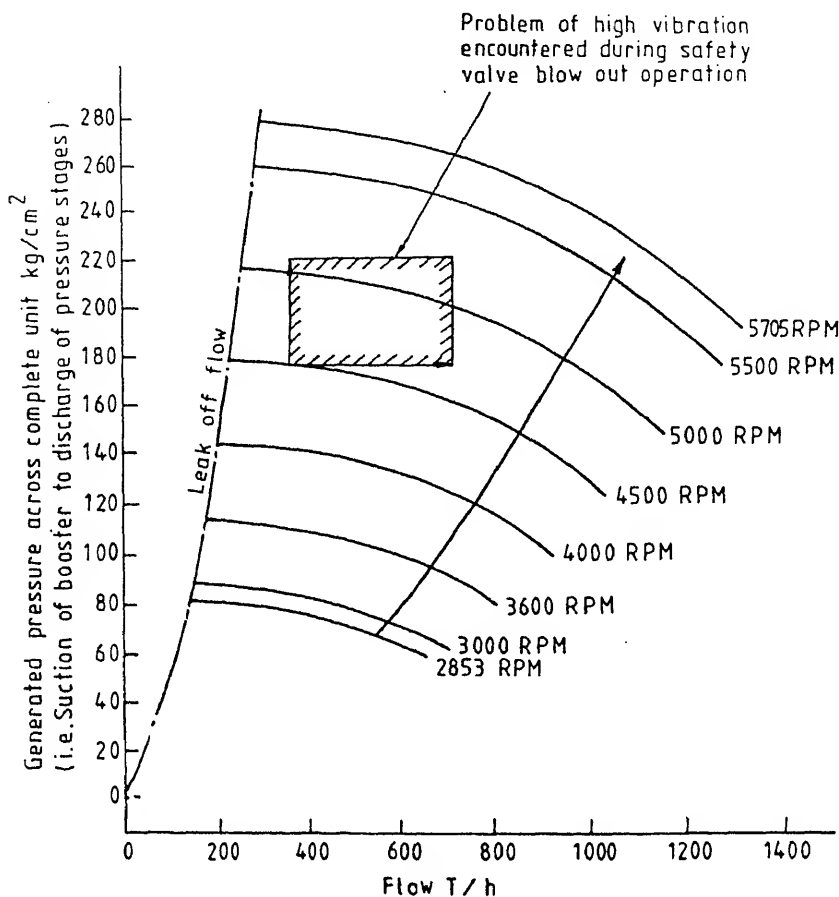
Figure 8.   Motor-driven BFP performance curve on cold water (s.g. 1.0).

Further, the natural frequencies of the piping structure were determined by the rap test, and were in the range of 2.5 to 3.0 Hz.

It was concluded that the severe vibrations of the interconnecting piping were due to resonance. Based on the experimental results and also on the experience of other research workers on centrifugal pumps as used in powerplants, the possibility of the following causes is mooted. A logical discussion is also included while deliberating on these causes.

(i) Presence of pressure pulsations even at non-severe conditions of operation has been identified but did not cause large piping vibrations because of non-coincidence of the excitation frequencies with the natural frequencies of the piping. Pressure pulsations are as a result of internal recirculation. It is well-known that centrigugal pumps are prone to internal recirculation at reduced flows (compared to rated flow). Consequences of internal recirculation are noisy operation and pressure pulsations (hydraulic surges) of low frequencies (1–8 Hz) causing pipeline vibration. It was noted that with increase in flow rate, pressure pulsation intensity was reduced, thus confirming the above reason for pressure pulsations. Pressure pulsations due to internal recirculation can be avoided by,
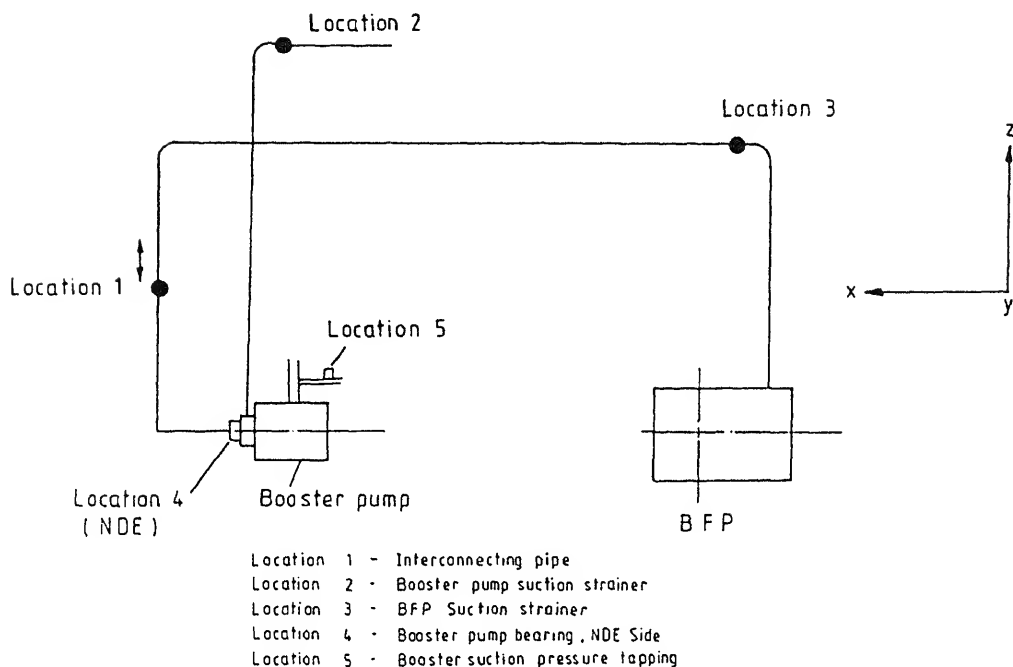
Figure 9. Booster pump, BFP and piping layout with measuring points location.

(a) reduction of pump inlet eye area by providing suitable inserts;

(b) running the pump at the rated flow capacity so as not to result in internal recirculation and deviating the excess flow at the delivery end through the recirculation pipeline, i.e. increasing the capacity of the recirculation valve and associated piping at the outlet of the BFP. This solution has worked very well in a somewhat similar situation (Pattabiraman *et al* 1986).

(ii) Referring to figure 8, severe vibrations of the piping have occurred at BFP discharge pressures of $176 \, \text{kg/cm}^2$ and flow rates of 350 t/h onwards. These conditions are needed to make the safety valve operative but at the discharge pressure of $168 \, \text{kg/cm}^2$ piping vibrations are reported to be of low level. It seems that floating of the safety

Table 4. Frequency components (in hertz) of vibration signatures.

| Location 1 | | | Location 2 | | | Location 3 | | | Location 4 | Location 5 |
V	T	A	V	T	A	V	T	A	H	V
—	—	1.64	—	1.7	—	—	—	—	2.0	1.9
2.4	2.4	2.4	—	2.4	2.5	2.9	2.6	2.6	2.8	—
3.4	3.3	3.3	—	—	—	3.4	3.4	—	—	—
5.7	5.8	5.8	6.1	6.0	6.0	5.8	5.8	6.0	6.0	5.9
6.9	—	7.1	—	—	—	—	—	7.3	6.6	—
8.1	8.5	8.5	—	—	—	8.5	8.5	8.5	8.4	—
9.5	9.7	9.5	9.2	9.2	9.0	9.4	9.4	9.7	9.4	9.6

Abbreviations: A – axial, H – horizontal, T – trans, V – vertical.

column from boiler to pump and pump to deaerator tank causing resonance of the piping system. Simpson & Tramschek (1974) have conducted a study on pulsations in power station feed-pump systems and reported that fluid column frequencies can be as low as 1.5 Hz, and any low frequency excitation may set in pressure fluctuations in the column.

8.3 Vibration problem at the non-driven end bearing of a generator of the gas based powerplant (Gupta & Yadava 1991)

Very high vibration level of $2 \times N$ frequency at the non-driven end (NDE) bearing of a generator was reported. It was required to identify the cause of this high vibration level.

Vibration measurements were carried out at the DE and NDE bearings of the generator in all the three directions and also for two load settings, viz. 61 and 82 MW. This was done to see whether the electrical load had any influence on the vibration level. Spectral analysis was performed on the recorded signatures, and some useful information from this analysis is presented in table 5.

In this table vibration levels in millimetres per second at frequencies $1 \times N, 2 \times N$ and $4 \times N$ (N being rotational speed of the rotor) are presented for loads of 61 and 82 MW. Dominance of the second harmonic vibration over the first harmonic vibration is clearly visible, particularly in the vertical direction at the NDE bearing, but the level is not affected much by the change in electrical load except at the NDE bearing in the horizontal direction. The frequency components which are greatly affected by the change of electrical load are underlined. These are the $4 \times N$ component in the vertical direction at both the DE and NDE bearings, and the $2 \times N$ component in the horizontal direction only at the NDE bearing. This information may help in tracing the cause of trouble by making use of the following knowledge-base.

The generator is a two-pole machine, so any defect producing a change in air gap between the pole and stator leads to an unbalanced magnetic pull, whose intensity increases with the flux, which in turn varies directly with the load. The unbalanced magnetic pull occurs twice in one rotation of the rotor, and gives rise to vibrations of $2 \times N$ frequency and

Table 5. Vibration levels at frequencies $1 \times N, 2 \times N$ and $4 \times N$ ($N = 3000$ rpm).

Location direction	Frequency	Vibration level (mm/s)			
		DE bearing load		NDE bearing load	
		61 MW	82 MW	61 MW	82 MW
Vertical	$1 \times N$	0.56	0.62	0.67	0.68
	$2 \times N$	3.55	4.12	8.42	8.52
	$4 \times N$	0.11	0.27	0.17	0.94
Axial	$1 \times N$	0.54	0.55	1.27	1.20
	$2 \times N$	3.13	3.16	2.99	2.79
	$4 \times N$	0.05	0.07	0.50	0.59
Horizontal	$1 \times N$	1.50	1.51	1.55	1.76
	$2 \times N$	5.44	5.19	0.13	0.35
	$4 \times N$	0.09	0.07	0.30	0.25

'multiples of this frequency in the stator body or bearing support. The magnitude of the higher harmonics depends on the shape of the unbalanced magnetic pull. Further, if there are two defects located 45° from the horizontal plane and on either side of the vertical axis, the net dominant unbalanced pull will be in the vertical direction and its frequency is $4 \times N$. The $4 \times N$ component in the horizontal direction gets moderated with the increase in load because of the opposing action. The enormous increase in the $2 \times N$ vibration level with the electrical load in the horizontal direction at NDE bearing suggests the presence of a defect in the stator at the level of its axis and near the NDE bearing. Similarly, the enormous increase in vibration level of the $4 \times N$ harmonic in the vertical direction both at DE and NDE bearings suggests the presence of two defects located 45° from the horizontal plane on either side of the vertical axis and situated somewhere along the length of the stator. It may be noted that the $4 \times N$ harmonic in the horizontal direction is somewhat moderated with the increase in electrical load.

It was recommended that the generator be opened to inspect the stator for any defect causing a change in the air gap, and on doing this it was found that the cover strips near the NDE bearing were absent. These were put back and the problem of high $2 \times N$ vibrations was eliminated.

9. Concluding remarks

Vibration is a very effective tool for diagnosing many mechanical defects. It however needs expert knowledge and experience for fault-tracing. When vibrations are used for condition monitoring, it is always advisable to start with simple instrumentation and before applying the technique to plant or machinery maintenance, the economical viability must be assessed. It is envisaged that in the near future the use of computers in handling the large amounts of monitored data and reducing them to usable information formats will expand. The tendency to use expert systems for fault diagnosis will also increse. There will be attempts to develop expert systems dedicated to a class of specific plants and machinery.

References

Collacott R A 1973 *Mechanical fault diagnosis* (London: Chapman and Hall)
Collacott R A 1977 *Fault diagnosis and condition monitoring* (London: Chapman and Hall)
Gupta K N 1986 Vibration monitoring – an overview. *Proc. Workshop on Diagnostic Maintenance,* New Delhi, pp 45–62
Gupta K N 1990 Vibration monitoring – state-of-the-art and future trends. *Proc. State-of-Art and Future Vision Seminar on Condition Monitoring,* New Delhi, pp II/1–16
Gupta K N, Yadava G S 1991 Vibration problem at NDE bearing of the generator at Anta Gas Project. Report submitted to NTPC, New Delhi
Gupta K N, Asnani N T, Nakra B C, Tandon N 1986 Cause analysis of dryer foundation vibrations at IPCL Complex, Baroda. *Proc. Workshop on Performance Assurance in Machinery,* Ahmedabad, pp MD 1–8

Gupta K N, Gupta K, Reddy P V 1988 Diagnosis of BFP inter-connecting piping vibration through signature analysis. *Proc. 17th National Convention on Maintenance Management for Higher Productivity, IIPE*, Jamshedpur, pp III/2/1–18

Pattabiraman J, Srinivasan R, Bhattacharjee D K 1986 Vibration studies on a suction piping of a boiler feed pump. *Symp. on Vibration Problems in Nuclear Power Stations*, Bombay

Simpson H C, Tramschek A B 1974 Pulsations in power station feed pump systems. *Proc. Inst. Mech. Engineers, Conf. on Component Interaction in Fluid Flow Systems*, London

Tranter J 1989 The fundamentals of, and the application of computers to condition monitoring and predictive maintenance. *Proc. Int. Congress on Condition Monitoring and Diagnostic Engineering (COMADEM 1989)*, pp 372–377

On the analysis of time-periodic nonlinear dynamical systems

S C SINHA

Nonlinear Systems Research Laboratory, Department of Mechanical Engineering, Auburn University, AL 36849, USA
e-mail: ssinha@eng.auburn.edu

Abstract.   In this study, a general technique for the analysis of time-period nonlinear dynamical systems is presented. The method is based on the fact that all quasilinear periodic systems can be replaced by similar systems whose linear parts are time invariant via the well-known Liapunov–Floquet (L–F) transformation. A general procedure for the computation of L–F transformation in terms of Chebyshev polynomials is outlined. Once the transformation has been applied, a periodic orbit in original coordinates has a fixed point representation in the transformed coordinates. The stability and bifurcation analysis of the transformed equations are studied by employing the *time-dependent normal form theory and time-dependent centre manifold reduction.* For the two examples considered, the three generic codimension-one bifurcations, viz, Hopf, flip and tangent, are analysed. The methodology is semi-analytic in nature and provides a quantitative measure of stability even under critical conditions. Unlike the perturbation or averaging techniques, this method is applicable even to those systems where the periodic term in the linear part does not contain a small parameter or a generating solution does not exist due to the absence of the time-invariant term in the linear part.

Keywords.   Nonlinear dynamical systems; Liapunov–Floquet transformation; centre manifold reduction; normal form theory; bifurcations; time-invariant forms.

1.   Introduction

The study of systems governed by a set of ordinary differential equations is of great importance in diverse branches of science and engineering. Numerous practical applications can be found in the areas of quantum mechanics, systems and controls and dynamic stability of structures under oscillating loads. In particular, such problems arise in the dynamics of rotating systems such as helicopter blades (Bramwell 1976; Johnson 1980) and rotor-bearing

nonlinear non-autonomous differential equations which contain explicit periodic functions of time. Periodic solutions of these equations physically represent steady-state operations under various conditions. The stability of these periodic solutions (or orbits) is determined by the equations of perturbed motion about the periodic solutions. In many situations, the linearized perturbed equations may be sufficient for the prediction of stability, and therefore the problem reduces to a set of linear ordinary differential equations with periodic coefficients. The same mathematical problem also arises in the study of nonlinear autonomous systems when the stability of a particular periodic solution needs to be investigated. Besides the stability issues, the linear control problems associated with rotating systems can also lead to the same type of equations. For example, the Individual-Blade-Control (IBC) technique, used in the control of a helicopter rotor-blade system, produces a set of equations with periodic coefficients (Kretz 1976; Mckillip 1985). Therefore the analysis of this special class of time-varying systems has been deemed extremely important.

Hill's method (Lindh & Likins 1970; Yakubovitch & Starzhinskii 1975), perturbation techniques (Stoker 1950; Nayfeh 1973) and Floquet's theory (Coddington & Levinson 1955; Floquet 1983) are some of the most commonly used mathematical methods in the analysis of such systems. It is well known that Hill's approach is not suitable for digital implementation, especially if one has to deal with a large-scale system. The perturbation methods have their own limitations due to the fact that they can only be applied to systems where the periodic coefficients can be expressed in terms of a small parameter. Therefore, Floquet analysis coupled with a numerical integration code has served as the main tool in various applications (Peters & Hohenemser 1971; Friedmann *et al* 1977; Gaonkar *et al* 1981; Sinha & Wu 1991; Wu & Sinha 1994). Floquet analysis is a powerful technique which can be easily implemented on a computer and holds most promise in the analysis of large-scale periodic systems. According to this technique, stability analysis requires the eigen-analysis of the 'Floquet Transition Matrix' (FTM) which is simply defined as the state transition matrix at the end of one period. Eigen-analysis and the evaluation of FTM are the two major computational problems encountered in the analysis of linear periodic systems of large dimensions. Since the FTM is nonsymmetric, the eigen-analysis problem is a difficult one. Some progress has been made in this direction (Cullum & Willoughby 1986; Gaonkar & Peters 1986), however, the computation of FTM for large-scale systems is still a challenging task. Most commonly a fourth or higher-order Runge–Kutta type numerical code has been used in a 'single pass' scheme for an efficient computation of the FTM (Friedmann *et al* 1977; Gaonkar *et al* 1981). Very recently, Sinha and associates (Sinha & Wu 1991; Joseph *et al* 1993; Sinha *et al* 1993; Wu & Sinha 1994) have developed a new technique for the analysis of large-scale periodic systems. It has been shown that the proposed technique is numerically several times faster than the standard codes and at the same time it can also be applied in the symbolic form (Sinha & Juneja 1991).

Although the linearized equations play an important role in the stability analyses, they fail to provide answers to many questions associated with the nonlinear periodic systems. Questions such as how does the solution depend on a control parameter? What kind of motion takes place after the loss of stability of the periodic state? Is it possible to identify the types of stability loss and critical values of the control parameters in each case etc? In order to answer such questions one must investigate the nonlinear equations of perturbed motion

including a bifurcation analysis of the periodic orbit. In many instances certain methods of nonlinear dynamics can be applied to obtain information of significant value. It is known (Arnold 1988) that the simplest loss of stability of a periodic orbit constitutes a degenerate problem of codimension 1. However, if a pair of complex multipliers simultaneously cross the unit circle of the complex plane, then we have the Hopf bifurcation of periodic orbits. The details are given by Guckenheimer & Holmes (1983) and Arnold (1988). For the case of a codimension 2 bifurcating periodic orbit, some qualitative mathematical results have also be obtained (Chow & Wang 1985). Qualitative analyses certainly provide good insight into the problems, but for engineering applications quantitative methods are indispensable.

Perturbation and averaging methods (Bogoliubov & Mitropolsky 1961; Nayfeh 1973; Sanders & Verhulst 1985) are suitable for relatively smaller systems and in general, their applications are limited to systems where the periodic terms as well as the nonlinearities can be expressed in terms of a suitable small parameter. Hopf bifurcation of Duffing's oscillator and nonlinear Mathieu's equations are discussed by Awreicewicz (1989) through an application of perturbation and harmonic balance methods. On the other hand, one can apply standard numerical techniques associated with boundary value problems to analyse the situation. These techniques are basically shooting methods and provide strategies for calculating branch points and new branches of bifurcating solutions. These methods have been exploited by several authors (Seydel 1981, 1987, 1988; Doedel & Kernévez 1986). Shooting methods are quite reliable but they are certainly not free from numerical instability difficulties. At the same time, when a system is non-autonomous, the trajectories can cross themselves and it is difficult to obtain a general structure of the motion through a purely numerical scheme. An attractive alternate method of analysis is provided by the technique called 'point mapping'. The idea was introduced by Poincaré (1899) and later developed by Birkhoff (1966), Arnold (1988) and Bernussou (1977). In this approach the continuous-time periodic system is reformulated as discrete-time events by defining a point mapping called the Poincaré map. Thus the original non-autonomous differential system is replaced by a set of difference equations which do not explicitly depend on time. In principle these are easier to analyse and simulate on a digital computer. However, one faces serious computational difficulties in application of this technique to real engineering problems even if the dimensions are small. In order to obtain the corresponding difference equations, one must construct an exact or approximate solution of a system of nonlinear differential equations. Exact solutions are only possible in very special cases, such as those of impulsive excitation problems discussed by Hsu and his associates (Hsu & Cheng 1973, 1974; Flashner & Hsu 1983; Hsu 1987). Since one must settle for an approximate representation of the point mapping, recent studies (Lukes 1982; Flashner & Hsu 1983; Guttalu & Flashner 1989, 1990) have suggested the use of Runge–Kutta type algorithm and perturbation technique for obtaining a truncated version of the Poincaré map. Following this approach one can discuss the bifurcation of periodic solutions to other possible periodic motions or to quasiperiodic and aperiodic solutions (Lindtner *et al* 1990).

One other viable approach is to use the Liapunov–Floquet theorem which allows transformation of the quasilinear periodic systems into a new set of similar equations whose linear parts are time-invariant. However, it is not a simple task to compute this transformation

matrix for a general periodic system. For certain special class of linear systems, it is possible to obtain the Liapunov–Floquet transformations as indicated by Lukes (1982). In order to determine such a transformation for a general periodic system, one must compute the STM as an explicit function of time. Recently we (Pandiyan *et al* 1993; Sinha & Joseph 1994; Sinha & Pandiyan 1994; Pandiyan & Sinha 1995) have been successful in developing a computational procedure through which the Liapunov–Floquet (L–F) transformation can be obtained in terms of Chebyshev polynomials which is suitable for algebraic manipulations. The inverse of the L–F transformation can also be computed by a similar procedure considering the adjoint system equation.

The development of a procedure for computing these transformation matrices has given a clear edge in dealing with a wide range of problems associated with periodically varying systems. In this paper, a quantitative analysis of nonlinear dynamical systems with periodic coefficients has been presented through an application of the Liapunov–Floquet (L–F) transformation. It is shown that the original quasilinear periodic system can be transformed to a dynamically similar form in which the linear part is time-invariant. The analysis of the transformed equations has been carried out through the use of *time-dependent normal form theory*. The solutions thus obtained are mapped back to the original coordinates by applying the inverse L–F transformations and compared with the numerical results obtained by a Runge–Kutta type algorithm. The method is also applicable to systems undergoing bifurcations. Such problems are referred to as 'critical cases' and have been studied through an application of the *centre manifold theory* (Malkin 1962; Carr 1981). For brevity, only codimension 1 bifurcations are considered. In order to demonstrate the effectiveness of the proposed analysis procedure, two examples have been studied in detail. The first example consists of a nonlinear Mathieu equation, the L–F transformation of which has been computed using the Chebyshev polynomials as described in § 3. The solutions of this example have been obtained in stable and centre manifolds for some typical sets of system parameters. It has been shown that the proposed technique is applicable to a wide class of problems including the situations where the generating solutions do not exist and/or the parameter multiplying the linear periodic terms are no longer small. It is also shown that in many cases it is possible to obtain approximate analytical solutions which compare extremely well with the numerical solutions. The results obtained by the traditional averaging method are also presented for comparison purposes.

As a second example, the bifurcation problem of a double inverted pendulum subjected to periodic loading is selected. The Hopf bifurcation in a double inverted pendulum subjected to a tangential static load has been studied by Sethna & Shapiro (1977) and thereafter many researchers have contributed on various bifurcation aspects of such an autonomous system. However, when the double pendulum is subjected to a periodic load, the system becomes non-autonomous. Periodic bifurcations of such a pendulum has been reported by Flashner & Hsu (1983) by the method of point mappings. In this paper, the dynamics of this four-dimensional system undergoing a single Hopf bifurcation or a single flip bifurcation is investigated in a two-dimensional centre manifold or a single-dimensional centre manifold, respectively by applying the time-dependent normal form theory and centre manifold reduction. The results of such analyses are verified by using numerical simulations.

2. Background

2.1 *Mathematical structure of periodic systems*

In general, many problems of mechanical systems can be reduced to a set of nonlinear ordinary differential equations of the form,

$$\dot{\mathbf{y}} = \mathbf{g}(y, \lambda, t), \tag{1}$$

where $\mathbf{y}$ is an n state vector, $g(\ )$ is a continuous nonlinear function of y and t and λ is a set of control parameters.

Let $\bar{y}(t)$ denote a known periodic solution (of period T) of (1) such that the perturbation $x(t)$ about the periodic motion $\bar{y}(t)$ can be defined as

$$y(t) = \bar{y}(t) + x(t). \tag{2}$$

Substituting (2) in (1) and expanding the right-hand side in Taylor series about $\bar{y}(t)$ yields

$$\dot{x}_i(t) = \left.\frac{\partial g_i}{\partial y_j}\right|_{\bar{y}(t)} x_j + \frac{1}{2}\left.\frac{\partial^2 g_i}{\partial y_j \partial y_k}\right|_{\bar{y}(t)} x_j x_k + \frac{1}{3!}\left.\frac{\partial^3 g_i}{\partial y_j \partial y_k \partial y_l}\right|_{\bar{y}(t)} x_j x_k x_l + \cdots. \tag{3}$$

The above equation may be rewritten as

$$\dot{\mathbf{x}} = \mathbf{A}(t, \lambda)x + \mathbf{f}_2(x, t, \lambda) + \mathbf{f}_3(x, t, \lambda) + \cdots + \mathbf{f}_k(x, t, \lambda) + \mathbf{O}(|x|^{k+1}, t) \tag{4a}$$

or

$$\dot{\mathbf{x}} = \mathbf{A}(t, \lambda)x + \mathbf{F}(x, t, \lambda) \tag{4b}$$

where $\mathbf{f}_k(\ )$ contain homogeneous monomials in x_i of order k. $\mathbf{A}(t, \lambda), \mathbf{f}_k(x, t, \lambda)$ and $\mathbf{F}(x, t, \lambda)$ (suitably defined in terms of $f_k(\ )$'s) are T periodic functions. The linear part of (4) is, of course, given by

$$\dot{\mathbf{x}} = \mathbf{A}(t, \lambda)x; \quad \mathbf{A}(t, \lambda) = \mathbf{A}(t + T, \lambda). \tag{5}$$

The stability and response of (5) can be discussed using the well-known Floquet theory.

We are interested in the analysis of (4) and determine its behaviour as the control parameter λ varies.

2.2 *Results from Floquet theory* (Floquet 1983; Coddington & Levinson 1955)

Theorem 1. *Each state transition matrix (STM), Φ of (5) can be written as the product of two $n \times n$-matrices as*

$$\Phi(t) = \mathbf{L}(t)e^{tC} \tag{6}$$

where $\mathbf{L}(t)$ is T-periodic and $\mathbf{C}$ is a constant $n \times n$ matrix. $\mathbf{L}(t)$ and $\mathbf{C}$, in general, are

Remark 1. There exists a complex matrix **C** such that

$$\mathbf{M} = e^{CT}, \tag{7}$$

where **M** can be expressed as

$$\mathbf{M} = \Phi^{-1}(o)\Phi(T). \tag{8}$$

Remark 2. μ_i, the eigenvalues of **M** are called the characteristic multipliers and the stability condition can be expressed as $|\mu_i| < 1, i = 1, 2, \ldots, n$.

COROLLARY 1
Each state transition matrix $\Phi(t)$ can also be factored as

$$\Phi(t) = \mathbf{Q}(t)e^{Rt} \tag{9}$$

*where the matrix $\mathbf{Q}(t)$ is real and periodic with period 2T and **R** is an appropriate real matrix.*

COROLLARY 2
The Liapunov–Floquet transformation

$$\mathbf{x}(t) = \mathbf{L}(t)z(t) \tag{10}$$

reduces the original time-varying system (5) to

$$\dot{z}(t) = \mathbf{C}z(t) \tag{11}$$

which is time-invariant.
  Moreover, the 2T-periodic transformation

$$x(t) = \mathbf{Q}(t)z(t) \tag{12}$$

produces a real representation given by

$$\dot{z}(t) = \mathbf{R}z(t) \tag{13}$$

3. Computation of L–F transformation matrix via Chebyshev polynomials

It has been shown by Sinha and associates (Sinha & Juneja 1991; Sinha & Wu 1991; Joseph *et al* 1993; Sinha *et al* 1993; Wu & Sinha 1994) that the STMs of linear periodic systems can be obtained in terms of the shifted Chebyshev polynomials of the first kind. The technique is efficient and since the STM is expressed as an explicit function of time t, it is suitable for algebraic manipulations as well. In fact, if the dimension is small, the STM can be expressed in a closed form as an explicit function of system parameters as shown by Sinha & Juneja (1991) for the case of Mathieu equation.

  In order to compute the L–F transformation matrix, $\mathbf{L(t)}$, one needs to find the STM $\Phi(t)$ associated with the linear system given by (5). If $\mathbf{A(t)}$ in (5) is commutative, then $\mathbf{L(t)}$ can be computed as

$$\mathbf{L}(t) = e^{\mathbf{B}_T(t)} \tag{14}$$

where $\mathbf{B}_T(t)$ can be obtained in a closed form as shown by Lukes (1982). For a general $\mathbf{A}(t)$, first, the Chebyshev polynomial expansion technique is used to compute the STM $\Phi(t)$. In this technique, the solution vector $\mathbf{x}(t)$ and the periodic matrix $\mathbf{A}(t)$ in equation (5) are expanded in terms of the shifted Chebyshev polynomials in the interval $[0, T]$ as shown below.

$$x_i(t) \approx \sum_{r=0}^{s-1} b_r^i s_r^*(t) \equiv \mathbf{s}^{*T}(t)\mathbf{b}^i, \quad i = 1, 2, \ldots, n \tag{15}$$

$$\mathbf{A}(t) \approx \sum_{r=0}^{s-1} d_r^{ij} s_r^*(t) \equiv \mathbf{s}^{*T}(t)\mathbf{d}^{ij}, \quad i, j = 1, 2, \ldots, n \tag{16}$$

where b_r^i are unknown expansion coefficients of $x_i(t)$, d_r^{ij} are known expansion coefficients of $A_{ij}(t)$ and $s_r^*(t)$ are the shifted Chebyshev polynomials of the first kind. For convenience in algebraic manipulation an $n \times nm$ Chebyshev polynomial matrix is defined as

$$\hat{\mathbf{S}}(t) = \mathbf{I} \otimes \mathbf{s}^{*T}(t), \tag{17}$$

where $\otimes$ represents the Kronecker product (Sinha & Wu 1991), and $\mathbf{I}$ is an $n \times n$ identity matrix. Using the definitions in (15), (16) and (17), $x(t)$ and $\mathbf{A}(t)$ can be rewritten as

$$\mathbf{x}(t) = \hat{\mathbf{S}}(t)\bar{\mathbf{b}}, \mathbf{A}(t) = \hat{\mathbf{S}}(t)\mathbf{D} \tag{18}$$

$$\mathbf{A}(t)\mathbf{x}(t) = \hat{\mathbf{S}}(t)\bar{\mathbf{Q}}\bar{\mathbf{b}} \tag{19}$$

where $\bar{\mathbf{b}} = \{\mathbf{b}^1\mathbf{b}^2\mathbf{b}^3 \cdots \mathbf{b}^n\}^T$ is an $nm \times 1$ vector, $\mathbf{D} = [\mathbf{d}^{i1}\mathbf{d}^{i2}\mathbf{d}^{i3} \cdots \mathbf{d}^{ij}]$, $\mathbf{ij} = 1, 2, 3, 4, \ldots n$, is an $nm \times n$ matrix and $\bar{\mathbf{Q}}$ is an $nm \times nm$ product operation matrix (for details see Sinha & Wu 1991).

Substituting equations (18) and (19) in the integral form of equation (5), the unknown constant $\mathbf{b}$ can be determined by a set of linear algebraic equations. Therefore, the solution vector $\mathbf{x}$ can be determined from equation (15). However, the computation of $\Phi(t)$ requires a set of solutions of (5) with n initial conditions: $\mathbf{x}_i(0) = (1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $(0, 0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 1)$. By defining the resulting set of Chebyshev coefficient vector $\bar{\mathbf{b}}_i' s$ of the n solutions in a matrix form, the STM can be written as (see Sinha & Juneja 1991; Sinha & Wu 1991; Joseph *et al* 1993; Sinha *et al* 1993; Wu & Sinha 1994),

$$\Phi(t) = \hat{\mathbf{S}}(t)\bar{\mathbf{B}}, \tag{20}$$

where $\bar{\mathbf{B}} = [\bar{\mathbf{b}}_1\bar{\mathbf{b}}_2\bar{\mathbf{b}}_3 \cdots \bar{\mathbf{b}}_n]$ and $\Phi(0) = \mathbf{I}$. It has to be noted that the STM is valid only for $0 \le t \le T$ since the shifted Chebyshev polynomials of the first kind are defined over the interval $[0, T]$. When $t > T$, the STM can be evaluated using Floquet theory (Yakubovitch & Starzinskii 1975) as

$$\Phi(t) = [\Phi(\varsigma)][\Phi(T)]^n; \quad t = \varsigma + nT, \quad \varsigma \epsilon [0, T], \quad n = 1, 2, \ldots, \epsilon. \tag{21}$$

Once $\Phi(t)$ is known, the T-periodic complex matrix $\mathbf{L}(t)$ or the $2T$-periodic real matrix $\mathbf{Q}(t)$ can be computed in the following way (Pandiyan *et al* 1993; Sinha & Joseph 1994; Sinhá & Pandiyan 1994; Pandiyan & Sinha 1995). Since $\Phi(0) = \mathbf{I}$, $\mathbf{L}(0) = \mathbf{L}T = \mathbf{I}$, the

where $\mathbf{C}$ is a $n \times n$ constant complex matrix. By performing an eigenanalysis on $\Phi(T)$, the matrix $\mathbf{C}$ can be computed easily. Then the T-periodic L–F transformation matrix is

$$\mathbf{L}(t) = \Phi(t)e^{-Ct}. \tag{23}$$

In order to evaluate the $2T$-periodic real L–F transformation matrix $\mathbf{Q}(T)$, first we note that (*cf* Yakubovitch & Starzinskii 1975),

$$\Phi(2T) = \Phi^2(T) = e^{CT}e^{C^*T} = e^{2RT}, \tag{24}$$

where $\mathbf{C}^*$ is the conjugate matrix of $\mathbf{C}$, the $n \times n$ constant real matrix $\mathbf{R} = [\mathbf{C} + \mathbf{C}^*]/2$ and the 2-T periodic L–F matrix can be represented as

$$\begin{aligned}
\mathbf{Q}(t) &= \Phi(t)e^{-Rt}; \ 0 \le t \le T; \\
\mathbf{Q}(\tau + T) &= \Phi(\tau)\mathbf{Q}(T)e^{-R\tau}; \ T \le (T + \tau) \le 2T; \ 0 \le \tau \le T.
\end{aligned} \tag{25}$$

It should be noted that $\mathbf{Q}(t) = \mathbf{Q}(t + 2T)$.

If one is interested in finding $\mathbf{L}^{-1}(t)$ or $\mathbf{Q}^{-1}(t)$, then there are two avenues. $\mathbf{L}(t)$ and $\mathbf{Q}(t)$ may possibly be inverted through the use of a symbolic software like MACSYMA/MATHEMATICA/MAPLE. However, this is neither realistic or even possible at this point in time. The other approach is to first find the STM $\Psi(t)$ of the adjoint system

$$\dot{w} = -\mathbf{A}^T(t)w, \tag{26}$$

and use the following relationship (*cf* Yakubovitch & Starzinskii 1975),

$$\Phi^{-1}(t) = \Psi^T(t). \tag{27}$$

The computation of $\Phi^{-1}(t)$ is critical in determining $\mathbf{L}^{-1}(t)$ or $\mathbf{Q}^{-1}(t)$. For example, $\mathbf{L}^{-1}(t)$ can be evaluated utilizing the properties of the adjoint system as shown below,

$$\mathbf{L}^{-1}(t) = [\Phi(t)e^{-Ct}]^{-1} = e^{Ct}\Phi^{-1}(t) = e^{Ct}\Psi^T(t). \tag{28}$$

Such an approximation of L–F transformations has been found to be extremely convergent (Pandiyan *et al* 1993; Sinha & Joseph 1994; Sinha & Pandiyan 1994; Pandiyan & Sinha 1995) and since it is periodic with period T, the elements $\mathbf{L}_{ij}(t)$ and $\mathbf{Q}_{ij}(t)$ have the truncated Fourier representations

$$\mathbf{L}_{ij}(t) \approx \sum_{n=-q}^{q} c_n \exp(i2\pi nt/T), \quad i = \sqrt{-1} \tag{29}$$

$$\mathbf{Q}_{ij}(t) \approx \frac{a_0}{2} + \sum_{n=1}^{q} a_n \cos\frac{\pi nt}{T} + \sum_{n=1}^{q} b_n \sin\frac{\pi nt}{T} \tag{30}$$

respectively.

Since complex matrix $\mathbf{L}(t)$ (or the real matrix $\mathbf{Q}(t)$) can be computed as a function of t, all algebraic manipulations involving this matrix can be done in symbolic form using MATHEMATICA or MACSYMA. $\mathbf{L}_{ij}^{-1}(t)$ and $\mathbf{Q}_{ij}^{-1}(t)$ have similar Fourier representations.

In this study only the real L–F transformation $\mathbf{Q}(t)$ has been used to make it more appealing to the engineering community. A 12- to 15-term Chebyshev expansion has been found to yield extremely accurate representation of $\mathbf{Q}(t)$. The accuracy of $\mathbf{Q}(t)$ is directly dependent upon the convergence and accuracy of the STM $\Phi(t)$ itself. A convergence study in the computation of $\Phi(t)$ has been reported by Joseph *et al* (1993).

4. Time-dependent normal forms and centre manifold reduction

4.1 Normal forms

The fact that $\mathbf{A}(t)$ is time dependent in (4), a direct application of *normal form theory* is not possible. Using the transformation

$$\mathbf{x}(t) = \mathbf{Q}(t)z(t), \tag{31}$$

(4) takes the form

$$\dot{z} = \mathbf{R}z + \mathbf{Q}^{-1}(t)\{\mathbf{f}_2(z, t) + \mathbf{f}_3(z, t)\}, \tag{32}$$

where $\mathbf{R}$ is an $n \times n$ constant matrix and nonlinear terms of order four and higher have been neglected, since for generic codimension 1 bifurcations, the fourth order terms do not affect the local stability behaviour (Arnold 1988). The form of (32) is amenable to direct application of the method of *time dependent normal forms* (TDNF) for equations with periodic coefficients as shown by Arnold (1988).

Equation (32) in its Jordan canonical form can be written as

$$\dot{y} = \mathbf{J}y + w_2(y, t) + w_3(y, t) \tag{33}$$

where $\mathbf{J}$ is the Jordan form of matrix $\mathbf{R}$ and $w_k(y, t)$ are $2T$-periodic functions and contain homogeneous monomials of y_i of order 2 and 3. Using a sequence of near identity transformations of the form

$$y = v + h_r(v, t), \tag{34}$$

where $h_r(v, t)$ is a formal power series in v of degree r $(r = 2, 3)$ with periodic coefficients having the principal period $2T$, (33) can be reduced to its simplest form

$$\dot{v} = \mathbf{J}v + w_2(v, t) + w_3(v, t). \tag{35}$$

It is important to note that the $w_2(v, t)$ and $w_3(v, t)$ contain only a finite number of Fourier harmonics. This is due to the fact that the solution of the resulting homological equation depends on the resonance condition relating the eigenvalues of $\mathbf{J}$ and the Fourier frequencies of $\mathbf{w}_r(v, t)$ (Arnold 1988). It should be pointed out that the solution of the time-dependent homological equation requires the solution of a large set of linear algebraic equations even for a 2×2 system. For example, if for such a system, the L–F transformation matrix $\mathbf{Q}(t)$ is represented by a fifteen-term complex Fourier expansion and let us say that the degree of the monomials $r = 3$, then one needs to solve (2×124) equations in blocks of 31.

4.2 Centre manifold reduction

In situations where some of the eigenvalues of $\mathbf{J}$ in (33) are critical, the stability of (33) can be discussed in the centre manifold via *time-periodic centre manifold theorems*. Application of the normal form procedure to the reduced set of equations in the centre manifold

following, a theorem due to Malkin (1962) has been utilized to develop a practical method for finding the centre manifold relations for the time-periodic systems.

Let us assume that (33) has n_1 eigenvalues that are critical and n_2 eigenvalues that have negative real parts. Therefore, (33) may be rewritten in the form

$$\begin{Bmatrix} \dot{y}_c \\ \dot{y}_s \end{Bmatrix} = \begin{bmatrix} \mathbf{J}_c & 0 \\ 0 & \mathbf{J}_s \end{bmatrix} \begin{Bmatrix} y_c \\ y_s \end{Bmatrix} + \begin{Bmatrix} w_{c2} \\ w_{s2} \end{Bmatrix} + \begin{Bmatrix} w_{c3} \\ w_{s3} \end{Bmatrix}, \tag{36}$$

where the subscripts c and s represent the critical and stable vectors, respectively. According to the *centre manifold theorem*, there exists a relation (Malkin 1962; Pandiyan & Sinha 1995)

$$y_s = h(y_c, t), \tag{37}$$

such that $h(y_c, t)$ is of the form,

$$h(y_c, t) = \sum B_s^{(m_1 \cdots m_{n_1})}(t) y_1^{m_1} \cdots y_{n_1}^{m_{n_1}}; \; m_1 + \cdots + m_{n_1} \geq 1, \tag{38}$$

where $B_s^{m_1 \cdots m_{n_1}}(t)$ are periodic coefficients with period $2T$. The relation y_s given by equation (32) can be obtained as the formal solutions of the equations (see Malkin 1962; Pandiyan & Sinha 1995)

$$\frac{\partial y_s}{\partial t} + \sum_{i=1}^{n_1} \frac{\partial y_s}{\partial y_c}(\mathbf{J}_s y_s + \mathbf{W}_s) = \mathbf{J}_c y_c + \mathbf{W}_c \tag{39}$$

where $\mathbf{W}_c = w_{c2} + w_{c3}$ and $\mathbf{W}_s = w_{s2} + w_{s3}$ are nonlinear vector monomials of the critical and stable states of the system, respectively. It is important to note that the resulting solutions will be meaningful only if the coefficients $B_s^{m_1 \cdots m_{n_1}}(t)$ are also periodic. Although there exists an infinite number of expansions similar to (38) which have finite coefficients and also satisfy (39), there is only one with periodic coefficients. This result was first reported by Malkin (1962).

As a result of substitution of (38) in (39), a set of differential equations in terms of the unknown coefficients $B_s^{m_1 \cdots m_{n_1}}(t)$ is obtained in a form

$$\frac{d B_s^{(m_1 \cdots m_{n_1})}}{dt} - \lambda_j B_s^{(m_1 \cdots m_{n_1})} = C_s^{(m_1 \cdots m_{n_1})} \tag{40}$$

where λ_j, $j = 1, 2, \ldots, n_2$ are the eigenvalues of the stable part of the system and $C_s^{(m_1 \cdots m_{n_1})}$ are the known integral rational functions of the periodic coefficients on the right hand side of (40). The coefficients $B_s^{(m_1 \cdots m_{n_1})}$ can be obtained by formally solving the above set of differential equations. For this purpose, $B_s^{(m_1 \cdots m_{n_1})}$ is assumed in the form of a finite Fourier expansions as

$$\mathbf{B}_s^{(m_1 \cdots m_{n_1})}(t) = a_0 + \sum_{n=1}^{l} a_n \cos\left(\frac{2\pi n t}{\hat{T}}\right) + b_n \sin\left(\frac{2\pi n t}{\hat{T}}\right) \tag{41}$$

where $\hat{T} = 2T$. Substituting (41) in (40) and equating like terms on both sides of the

obtained. The constants a_n and b_n can be computed by solving these algebraic equations and therefore the coefficients $B_s^{(m_1 \cdots m_{n_1})}$ can be determined in the form of (41). Substitution of (32) in (36) clearly decouples the stable and critical states and, hence, the problem reduces to the investigation of stability of an n_1 dimensional system in the centre manifold. The resulting system of n_1 periodic equations is of the form

$$\dot{\mathbf{y}}_c = \mathbf{J}_c \mathbf{y}_c + \mathbf{W}_c^* \tag{42}$$

where vector $\mathbf{W}_c^*$ contains nonlinear monomials which are functions of $\mathbf{y}_c$ only.

5. Applications

To demonstrate the applicability and effectiveness of the suggested approach, two examples are considered. As a first example, a nonlinear Mathieu equation is considered for which the L–F transformation has been obtained using the computational algorithm discussed in § 3. Although this is a simple example, it brings out the key points clearly and shows the superiority of the proposed methods over the classical methods such as averaging, perturbation, etc. Whereas the results of the proposed technique, based on L–F transformation and normal forms, provide reasonably good solutions even for moderately large parameters multiplying the nonlinear terms, the traditional averaging procedure is applicable only when the parameters multiplying the periodic terms and the nonlinear terms are both small.

In the second example, the dynamic behavior of a time-periodic double pendulum undergoing various bifurcations is examined. In particular the critical dynamics under secondary Hopf and flip bifurcations are studied in detail.

5.1 *Example 1. Mathieu equation with cubic nonlinearity*

Consider the Mathieu equation with cubic nonlinearity in the form,

$$\ddot{x} + \delta \dot{x} + (\alpha + \beta \cos \omega t)x + \epsilon x^3 = 0, \tag{43}$$

where δ, α, β, ω and ϵ are the parameters of the system. In the state space form the above equation is rewritten as

$$\begin{Bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -(\alpha + \beta \cos \omega t) & -\delta \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} + \begin{Bmatrix} 0 \\ -\epsilon x_1^3 \end{Bmatrix} \tag{44}$$

where $\{\dot{x}_1, \dot{x}_2\}^T = \{x, \dot{x}\}^T$. Following the steps described in § 3, the 2-T periodic real L–F transformation matrix $\mathbf{Q}(t)$ can be computed for a given parameter set. Applying the transformation $\mathbf{x} = \mathbf{Q}(t)\mathbf{z}$, (44) is transformed to

$$\begin{Bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{Bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \begin{Bmatrix} z_1 \\ z_2 \end{Bmatrix} + \mathbf{Q}^{-1}(t) \begin{Bmatrix} 0 \\ -\epsilon (Q_{11}z_1 + Q_{12}z_2)^3 \end{Bmatrix}, \tag{45}$$

where R_{ii} are the elements of the real matrix $\mathbf{R}$ and Q_{ii} are the elements of the L–F trans-

(45) can be written in the canonical form as

$$
\left\{\begin{array}{c} \dot{y}_1 \\ \dot{y}_2 \end{array}\right\} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \left\{\begin{array}{c} y_1 \\ y_2 \end{array}\right\}
$$
$$
+ \epsilon \left\{\begin{array}{c} f_{11}(t,\tau)y_1^3 + f_{12}(t,\tau)y_1^2 y_2 + f_{13}(t,\tau)y_1 y_2^2 + f_{14}(t,\tau)y_2^3 \\ f_{21}(t,\tau)y_1^3 + f_{22}(t,\tau)y_1^2 y_2 + f_{23}(t,\tau)y_1 y_2^2 + f_{24}(t,\tau)y_2^3 \end{array}\right\} \tag{46}
$$

where $\tau = 2T$, $\Lambda = \{\lambda_1, \lambda_2\}$ are the eigenvalues of $\mathbf{R}$ and the periodic coefficients $f_{ij}(t,\tau); i, j = 1, \ldots, 4$ are expressed as

$$
f_{ij}(t,\tau) = a_0^{i,j} + \sum_{n=1}^{l} a_n^{i,j} \cos\left(\frac{2\pi n t}{\tau}\right) + \sum_{n=1}^{l} b_n^{i,j} \sin\left(\frac{2\pi n t}{\tau}\right). \tag{47}
$$

After experimenting with various sets of system parameters, it was observed that $l = 15$ provided quite accurate representations of functions $f_{ij}(t,\tau)$. This has been reported earlier by Pandiyan *et al* (1993). It is also consistent with the number of Fourier terms taken in the representation of the L–F transformation $\mathbf{Q}(t)$.

In order to obtain a solution of equation (46) using TDNF, consider a near-identity nonlinear transformation

$$
y_1 = u + g_{11}(t,\tau)u^3 + g_{12}(t,\tau)u^2 v + g_{13}(t,\tau)uv^2 + g_{14}(t,\tau)v^3,
$$
$$
y_2 = v + g_{21}(t,\tau)u^3 + g_{22}(t,\tau)uv^2 + g_{23}(t,\tau)uv^2 + g_{24}(t,\tau)v^3, \tag{48}
$$

where the periodic coefficients $g_{ij}(t,\tau); i, j = 1, \ldots, 4$ are once again of the form given by (47) but with unknown constants $\tilde{a}_n$ and $\tilde{b}_n$. Substituting (48) in (46) and solving the resulting homological equation as described earlier, the unknown constants $\tilde{a}_n$ and $\tilde{b}_n$ can be evaluated. In situations when there is no resonance, the Fourier series assumed for $g_{ij}(t,\tau)$ and its derivative are found to be convergent (Arnold 1988). On the other hand, if resonance takes place, the unknown constants of the corresponding periodic coefficient cannot be determined and some nonlinear terms remain even after the normal form reduction. As long as there is no resonance, all nonlinear terms are eliminated and the reduced normal form is just the linear part of (46). Therefore, the solution of the nonlinear Mathieu equation in the original coordinates can be obtained by substituting back all the intermediate transformations. Even when some of the nonlinear terms remain due to resonance, the resulting equation can still be used to provide many useful conclusions about the stability and dynamical behaviour of the system. Such procedures are described by Bruno (1989) and Hale & Kocak (1991).

In order to obtain an approximate solution of (46) via the *time independent normal form* (TINF) theory, variations of the periodic coefficients of nonlinear terms are neglected in comparison with their predominant means. This approximation results in an equation of the form

$$
\left\{\begin{array}{c} \dot{\tilde{y}}_1 \\ \dot{\tilde{y}}_2 \end{array}\right\} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \left\{\begin{array}{c} \tilde{y}_1 \\ \tilde{y}_2 \end{array}\right\} + \left\{\begin{array}{c} a_3 \tilde{y}_1^3 + a_4 \tilde{y}_1^2 \tilde{y}_2 + a_5 \tilde{y}_1 \tilde{y}_2^2 + a_6 \tilde{y}_2^3 \\ b_3 \tilde{y}_1^3 + b_4 \tilde{y}_1^2 \tilde{y}_2 + b_5 \tilde{y}_1 \tilde{y}_2^2 + b_6 \tilde{y}_2^3 \end{array}\right\}, \tag{49}
$$

ere $\tilde{y}_1$ and $\tilde{y}_2$ are the approximate states and $a_j, b_j, j = 3, \ldots, 6$ are the means of the
iodic coefficients of (47). To apply the TINF theory, consider a nonlinear near-identity
isformation of the form

$$
\begin{aligned}
\tilde{y}_1 &= u + p_1 u^3 + p_2 u^2 v + p_3 u v^2 + p_4 v^3, \\
\tilde{y}_2 &= v + q_1 u^3 + q_2 u^2 v + q_3 u v^2 + q_4 v^3,
\end{aligned}
\tag{50}
$$

ere p_i and q_i are unknown constants. Substituting (50) in (49) and solving the resulting
onomous homological equation, (49) can be reduced to a linear form in most of the
nations except when resonances due to nonlinearity occur. The approximate solutions
this case can also be obtained in a fashion similar to the procedure discussed above. It
lso observed that as long as the system has eigenvalues that are distinct with negative
positive real parts, one could completely reduce the system to a linear form.
At this point an application of the traditional averaging procedure to (43) is briefly
cussed. Assume a generating solution of the form (*cf* Sanders & Verhulst 1985),

$$
x(t) = z_1(t) \cos \omega_0 t + \frac{z_2(t)}{\omega_0} \sin \omega_0 t,
\tag{51}
$$

here $\omega_0^2 = \alpha$ and $z_i, i = 1, 2$ are the slowly varying coefficients of the solution. Using this
lution in (43) and averaging over the principal period $2\pi/\omega$, one gets a set of quasilinear
tonomous differential equations with cubic nonlinearity in the averaged coefficients $\tilde{z}$ as

$$
\begin{Bmatrix} \dot{\tilde{z}}_1 \\ \dot{\tilde{z}}_2 \end{Bmatrix} = \begin{bmatrix} R_{11}^0 & R_{12}^0 \\ R_{21}^0 & R_{22}^0 \end{bmatrix} \begin{Bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{Bmatrix} + \epsilon \begin{Bmatrix} \bar{a}_3 \tilde{z}_1^3 + \bar{a}_4 \tilde{z}_1^2 \tilde{z}_2 + \bar{a}_5 \tilde{z}_1 \tilde{z}_2^2 + \bar{a}_6 \tilde{z}_2^3 \\ \bar{b}_3 \tilde{z}_1^3 + \bar{b}_4 \tilde{z}_1^2 \tilde{z}_2 + \bar{b}_5 \tilde{z}_1 \tilde{z}_2^2 + \bar{b}_6 \tilde{z}_2^3 \end{Bmatrix},
\tag{52}
$$

here $\bar{a}_i$ and $\bar{b}_i, i = 3, \ldots, 6$ are the constants depending on the parameters of the system
d R_{ij}^0 are the elements of the constant matrix $\mathbf{R}^0$. However, the eigenvalues of the matrix
^{0} will be quite different from those appearing in (49). The solution of the above equation
n be discussed via the TINF theory.

2 Case studies

the following, several case studies are presented by selecting various sets of parameters
β, δ and ϵ in the Mathieu equation given by (43). The parameter ω is selected as 2π in
l cases of this study.

ase 1. Trajectories in stable manifold ($\alpha \neq 0$): *Parameter set 1:* $\alpha = 10.0$; $\beta = 0.2$;
$= 1.8974$; $\epsilon = 0.3$; Note that the parameters β and ϵ multiplying the periodic and
e nonlinear terms respectively are selected small. The suggested approaches, including
e *traditional averaging*, are applied to this set and the results, along with the numerical
lution, are presented in figure 1. It is understandable that all the methods predict the
ehaviour of the system correctly due to the smallness of the parameters β as well as ϵ.
or brevity, the coefficients of near-identity transformations for the normal forms are not
corded in this paper.
arameter set 2: $\alpha = 0.5$; $\beta = 4.0$; $\delta = 0.4243$; $\epsilon = 0.3$; For this case, β, the parameter
ultiplying the periodic term, is selected to be 8 times larger than α. From figure 2,
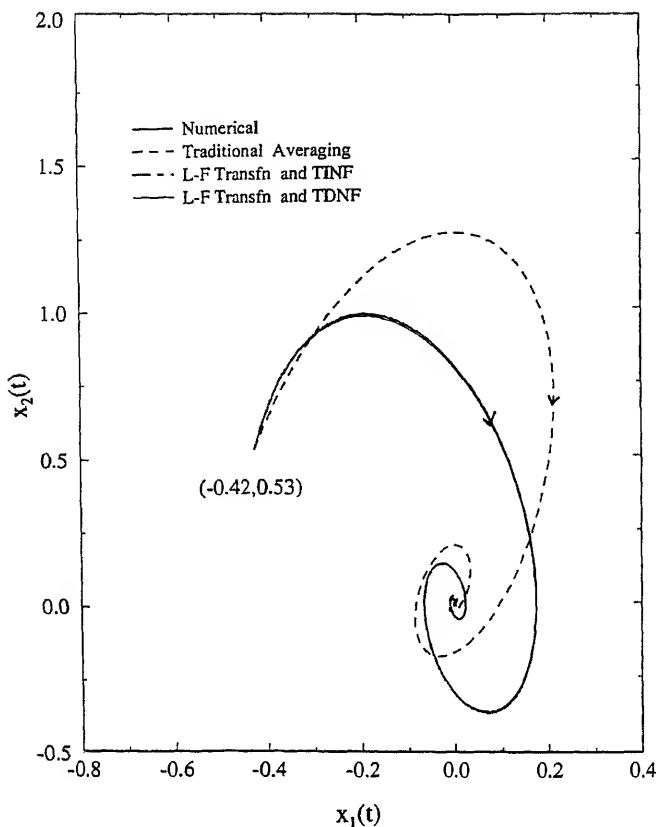
Figure 1. Comparison of solutions of Mathieu's equation with cubic nonlinearity: $\alpha = 10$, $\beta = 0.2$, $\delta = 1.8974$, $\epsilon = 0.3$.

it is observed that except for the averaging method, all other techniques predict similar trajectories which finally approach the fixed point $(0, 0)$.

Case 2. Equation without generating solution ($\alpha = 0$): Note that when $\alpha = 0$ the fundamental frequency ω_0 of the autonomous part of (43) is zero and without a generating solution the averaging approach cannot be applied.

Parameter set: $\alpha = 0$; $\beta = 4.0$; $\delta = 0.4243$; $\epsilon = 3.0$; This set shows that the behaviour of the system is well predicted by the suggested techniques when compared with the numerical solution even though the nonlinearity parameter ϵ is 10 times larger than the value used in set 1. The comparison is shown in figure 3.

The coefficients of the near-identity transformations for the TDNF as well as TINF methods are once again omitted for brevity.

Case 3. The centre manifold case: The application of the suggested techniques to the special case when the resonance condition prevails due to the presence of a pair of purely imaginary roots is shown in this part. For this case, by applying normal forms, all the non-resonant terms are annihilated, however, the homological equation corresponding to the resonant terms cannot be resolved. Therefore, some of the nonlinear terms of third
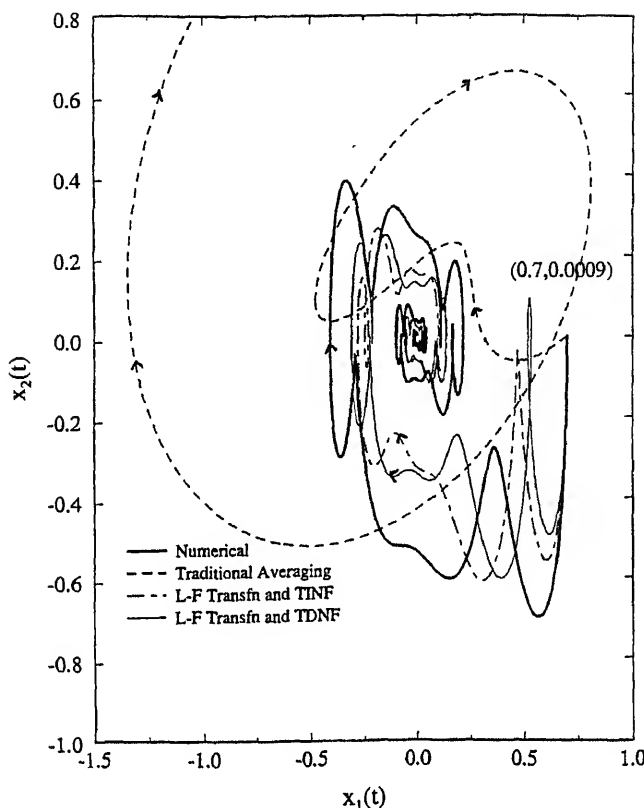
Figure 2. Comparison of solutions of Mathieu's equation with cubic nonlinearity: $\alpha = 0.5$, $\beta = 4.0$, $\delta = 0.4243$, $\epsilon = 0.3$.

degree stay in the reduced equation and hence, in general, a closed form solution is not possible. Under such circumstances it may be possible to neglect the periodic variations of the nonlinear coefficients and still retain the stability behaviour of the system. In the following, a case study is provided to demonstrate the applications in similar situations. *Parameter set:* $\alpha = 4.0$; $\beta = 7.468$; $\delta = 0$; $\epsilon = 0.3$; For this set of parameters, applications of the L–F transformation, the Jordan canonical transformation and the near-identity transformation (*cf* (48)) to (46) results in an equation of the form,

$$\begin{Bmatrix} \dot{u} \\ \dot{v} \end{Bmatrix} = \begin{bmatrix} -i\omega_0 & 0 \\ 0 & i\omega_0 \end{bmatrix} \cdot \begin{Bmatrix} u \\ v \end{Bmatrix} + \epsilon \begin{Bmatrix} f_{12}(t,\tau)u^2 v \\ f_{23}(t,\tau)uv^2 \end{Bmatrix}, \tag{53}$$

where, $f_{12}(t,\tau)$ and $f_{23}(t,\tau)$ are complexified periodic functions corresponding to the resonant terms and $\pm i\omega_0 = \pm i0.816$ are the eigenvalues of the system. Multiplying $\dot{u}$ by v and $\dot{v}$ by u and adding, a linear differential equation in (uv) can be obtained. Therefore, the analytical solution of (53) can be found in (uv) which is of the form

$$uv = -1 \bigg/ \int_0^t a(\chi,\tau)\,\mathrm{d}\chi, \tag{54}$$

where $a(t,\tau)$ is a complex Fourier function. The differential equation (53) can be decoupled by substituting the solution (54) into (53) and the resulting linear differential equation in
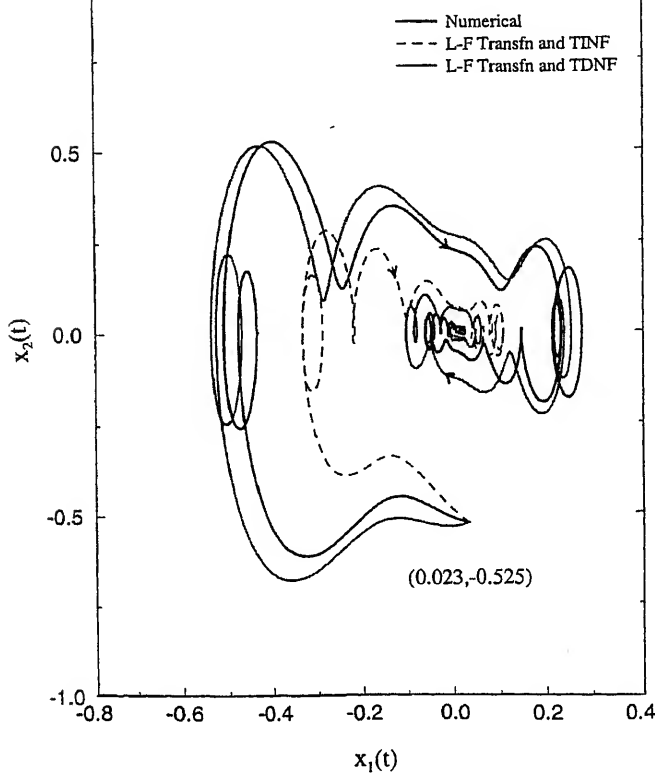
Figure 3. Comparison of solutions of Mathieu's equation with cubic nonlinearity: $\alpha = 0$, $\beta = 4.0$, $\delta = 0.18974$, $\epsilon = 0.3$.

u can be shown to be of the form,

$$\dot{u} = \left[-i\omega_0 + \epsilon f_{12}(t, \tau) \middle/ \left(-\int_0^t a(\chi, \tau) \, d\chi \right) \right] u,$$

$$\dot{u} = \left[-i\omega_0 + \epsilon \sum_{n=-q}^{q} c_n \exp(i2\pi nt/\tau) \right] u, \tag{55}$$

where c_n are complex constants of Fourier expansion of the periodic terms and hence

$$u = \exp\left[\left(-i\omega_0 + \epsilon \sum_{n=-q}^{q} c_n \exp(i2\pi nt/\tau) \right) t \right] u_0. \tag{56}$$

In a similar manner, v can also be computed. It should be noted that the stability of these solutions entirely depend on the real part of the constant coefficient of the time-varying function $c_n \exp(i2\pi nt/\tau)$. The solution is stable and/or unstable depending on whether the constant has negative real part or positive real part, respectively. This observation has also been made by Rosenblat & Cohen (1980, 1981) following an entirely different approach. When the real part is zero, the solutions are closed orbits and behave like limit cycles. A typical result is shown in figure 4.
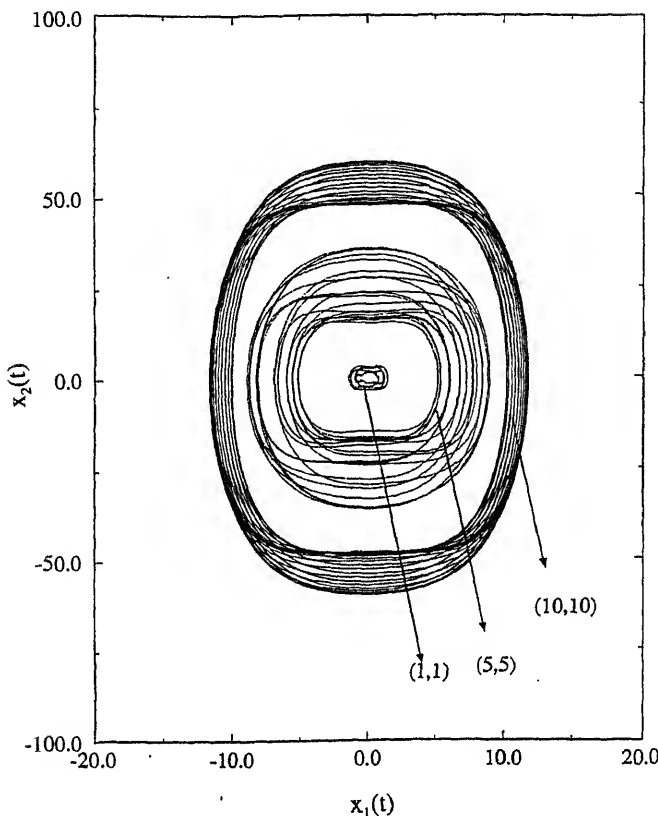
Figure 4. Phase plots of Mathieu's equation in centre manifold: $\alpha = 4.0$, $\beta = 7.468$, $\delta = 0$, $\epsilon = 0.3$.

Example 2. Double inverted pendulum with parametric excitation

In this example, the bifurcation of a two-mass inverted pendulum subjected to non-conservative periodic load is discussed. The nonlinear equations of motion of the system are of the form (Jin & Matsuzaki 1988; Pandiyan & Sinha 1995)

$$
\begin{aligned}
\ddot{\phi}_1 = & -0.5(B_1 + 2B_2)\dot{\phi}_1 + B_2\dot{\phi}_2 + 0.5\bar{k}(\bar{p} - 3)\phi_1 \\
& + 0.5\bar{k}(2 - \bar{p})\phi_2 - 0.5(\dot{\phi}_1^2 + \dot{\phi}_2^2)(\phi_1 - \phi_2) \\
& - (\bar{p}/\bar{k}/12)\{(\phi_1 - \gamma\phi_2)^3 - (1 - \gamma)^3\phi_2^3\} - ((\phi_1 - \phi_2)^2/4) \\
& \times \{\bar{k}(\bar{p} - 4)\phi_1 + \bar{k}(3 + \bar{p}(\gamma - 2))\phi_2 - (B_1 + 3B_2)\dot{\phi}_1 + 3B_2\dot{\phi}_2\},
\end{aligned}
\tag{57}
$$

$$
\begin{aligned}
\ddot{\phi}_2 = & \, 0.5(B_1 + 4B_2)\dot{\phi}_1 - 2B_2\dot{\phi}_2 + 0.5(5 - \bar{p})\bar{k}\phi_1 \\
& + \{(\bar{p}(1.5 - \gamma) - 2)\bar{k}\}\phi_2 + 0.5(\phi_1 - \phi_2)(3\dot{\phi}_1^2 + \dot{\phi}_2^2) \\
& + (\bar{p}\bar{k}/12)\{(\phi_1 - \gamma\phi_2)^3 - 3(1 - \gamma)^3\phi_2^3\} + ((\phi_1 - \phi_2)^2/4) \\
& \times \{\bar{k}(2\bar{p} - 7)\phi_1 + \bar{k}(5 + \bar{p}(\gamma - 3))\phi_2 - (2B_1 + 5B_2)\dot{\phi}_1 + 5B_2\dot{\phi}_2\},
\end{aligned}
\tag{58}
$$

where m is the mass, l is the length of the links of the pendulum, ϕ_1 and ϕ_2 the displacement angles, $\dot{\phi}_1$ and $\dot{\phi}_2$ are the corresponding rates, γ ($0 \leq \gamma \leq 1$) is the load-direction parameter and $P = P_1 + P_2 \cos \omega \tau$. Other symbols appearing in (57) and (58) are defined as b_1 & b_2 = damping parameters, $B_1 = b_1/ml^2$, $B_2 = b_2/ml^2$, $\bar{p} = Pl/ml^2$, $\bar{k} = k/ml^2$

k = stiffness parameter, P_1 = magnitude of static load, P_2 = amplitude of the dynamic periodic load. Equations (57) and (58) are rewritten in the state-space form as

$$\begin{Bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{Bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5\bar{k}(\bar{p}-3) & 0.5\bar{k}(2-\bar{p}) & -0.5(B_1+2B_2) & B_2 \\ 0.5\bar{k}(5-\bar{p}) & \bar{k}[\bar{p}(1.5-\gamma)-2] & 0.5(B_1+4B_2) & -2B_2 \end{bmatrix}$$

$$(59)$$

$$\times \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0 \\ [-0.5(y_3^2+y_4^2)(y_1-y_2)-(\bar{p}\bar{k})[(y_1-\gamma y_2)^3 \\ -(1-\gamma)^3 y_2^3]/12 - 0.25(y_1-y_2)^2[\bar{k}(\bar{p}-4)y_1 \\ +\bar{k}(3+\bar{p}(\gamma-2))y_2 - (B_1+3B_2)y_3+3B_2y_4]] \\ [0.5(y_1-y_2)(3y_3^2+y_4^2)+\bar{p}\bar{k}[(y_1-\gamma y_2)^3 \\ -3(1-\gamma)^3 y_2^3]/12+0.25(y_1-y_2)^2[\bar{k}(2\bar{p}-7)y_1 \\ +\bar{k}(5+\bar{p}(\gamma-3))y_2 - (2B_1+5B_2)y_3+5B_2y_4]] \end{Bmatrix}$$

where $\{y_1, y_2, y_3, y_4\} = \{\phi_1, \phi_2, \dot{\phi}_1, \dot{\phi}_2\}$. In the following, the dynamics of a primary single Hopf and a single flip bifurcations of the above 4-dimensional system is discussed via centre manifold principle by reducing the problem to a two and a single dimension respectively.

(i) *Hopf bifurcation* – For the parameter set, $\bar{k} = 2.0$, $B_1 = B_2 = 0.016$, $P_1 = 0.5$, $P_2 = 0.966$, $\gamma = 0.8$, $\omega = 1.0$, (46) yields a pair of complex Floquet multipliers with modulus one which corresponds to a single Hopf bifurcation. After normalizing the time with $\omega\tau = 2\pi t$, the L–F transformation corresponding to (59) is computed. The application of this transformation to (59) leads to the following dynamically equivalent Jordan canonical form.

$$\begin{Bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{Bmatrix} =$$

$$\begin{bmatrix} -0.17791+1.14613\,i & 0 & 0 & 0 \\ 0 & -0.17791-1.14613\,i & 0 & 0 \\ 0 & 0 & 0.3391\,i & 0 \\ 0 & 0 & 0 & -0.3391\,i \end{bmatrix}$$

$$(60)$$

$$\times \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} + \begin{Bmatrix} \sum a_q(t)y_1^{m_1}y_2^{m_2}y_3^{m_3}y_4^{m_4} \\ \sum b_q(t)y_1^{m_1}y_2^{m_2}y_3^{m_3}y_4^{m_4} \\ \sum c_q(t)y_1^{m_1}y_2^{m_2}y_3^{m_3}y_4^{m_4} \\ \sum d_q(t)y_1^{m_1}y_2^{m_2}y_3^{m_3}y_4^{m_4} \end{Bmatrix}, \quad \sum_{i=1}^{4} m_i = 3$$

where $a_q(t)$, $b_q(t)$, $c_q(t)$ and $d_q(t)$ are the complex vector periodic coefficients consisting of 31 elements (this corresponds to the number of Fourier terms taken in the expansion of

L–F transformation matrix) with period $2T$ corresponding to all possible monomials of order 3 in y_1, y_2, y_3 & y_4. Note that two of the eigenvalues in (60) are purely imaginary which is to be expected in this case. The centre manifold relations for this problem are assumed in the form

$$y_1 = B_{11}^3(t)y_3^3 + B_{12}^3(t)y_3^2y_4 + B_{13}^3(t)y_3y_4^2 + B_{14}^3(t)y_4^3,$$
$$y_2 = B_{21}^3(t)y_3^3 + B_{22}^3(t)y_3^2y_4 + B_{23}^3(t)y_3y_4^2 + B_{24}^3(t)y_4^3 \tag{61}$$

such that $B_{ij}^3(t)$, $i = 1, 2$ & $j = 1, 2, 3, 4$ are unknown periodic coefficients with period $2T$. Note that in the above equation the states corresponding to stable eigenvalues are expressed in terms of the states corresponding to the critical eigenvalues.

Substituting (61) in (60), eight ordinary differential equations in $B_{ij}^3(t)$, $i = 1, 2$ & $j = 1, 2, 3, 4$, similar to (40) are obtained. The periodic coefficients appearing on the right hand side of these differential equations are nothing but the known periodic coefficients corresponding to the cubic nonlinear terms appearing in (60). The unknown periodic coefficients $B_{ij}^3(t)$, $i = 1, 2$ & $j = 1, 2, 3, 4$ can be obtained by formally solving these differential equations. In order to obtain a particular solution, $B_{ij}^3(t)$ is assumed in the form of (41) with unknown constant coefficients and like terms on both sides of the equations are equated to obtain a set of linear algebraic equations in terms of the unknowns a_n and b_n. The computation of all the unknowns of the $B_{ij}^3(t)$'s requires the solution of a set of 8×31 linear algebraic equations. These algebraic equations can be solved such that each of the $B_{ij}^3(t)$, $i = 1, 2$ & $j = 1, 2, 3, 4$ can be obtained as Fourier series expansions. Noting that the problem under consideration consists of only cubic nonlinearities, it is not necessary to solve for all the periodic coefficients $B_{ij}^3(t)$, $i = 1, 2$ & $j = 1, 2, 3, 4$ in the centre manifold relation. Instead, it suffices to compute only one coefficient per relation in (61). This simplification does not affect the final outcome of the result, since the centre manifold relations result in nonlinearities which are of powers greater than three and does not influence the stability characteristics. Therefore, for this case, only coefficients $B_{11}^3(t)$ and $B_{21}^3(t)$ are computed.

Substitution of centre manifold relations (61) in (60) results in differential equations for the critical states y_3 and y_4 which contain nonlinearities of cubic and higher orders. Since the higher order terms do not affect the stability characteristics, the terms of order higher than 3 are neglected. The equations thus obtained are similar to (46) representing a Hopf bifurcation behaviour in a two-dimensional system with cubic nonlinearity. Following the procedure outlined earlier in the study of Hopf bifurcation of a single degree of freedom system, the application of time-dependent normal forms to these equations provide a simplified nonlinear equation similar to (53). The behaviour of the fixed point of the resulting equation is found to be a centre by employing similar methods to those outlined by Pandiyan (1994). For brevity, the calculations are not reported here. On the basis of the arguments presented for example 1, the motion resulting from the Hopf bifurcation is quasi-periodic and bounded. It can readily be seen that the Poincaré plots provided in figures 5a and 5b also confirm this result.
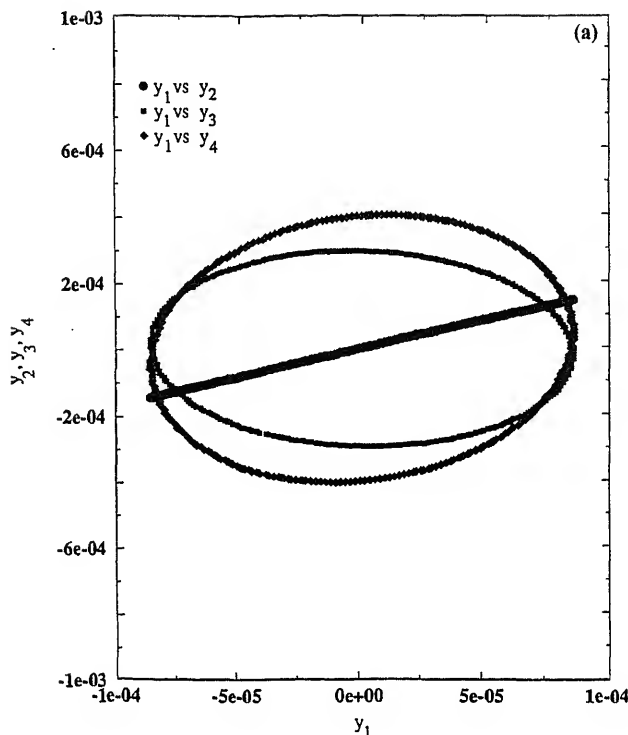
Figure 5.   (a) Poincaré plots – Double inverted pendulum under Hopf bifurcation.

$-0.999938 (\approx -1)$ and the system undergoes a flip bifurcation. After the transformation, the following Jordan canonical form is obtained

$$
\begin{Bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{Bmatrix} \approx \begin{bmatrix} -0.1685 + 0.9557\,i & 0 & 0 & 0 \\ 0 & -0.1685 - 0.9557\,i & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.0137 \end{bmatrix}
$$

(62)

$$
\times \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} + \begin{Bmatrix} \sum \mathbf{a}_q(t) y_1^{m_1} y_2^{m_2} y_3^{m_3} y_4^{m_4} \\ \sum \mathbf{b}_q(t) y_1^{m_1} y_2^{m_2} y_3^{m_3} y_4^{m_4} \\ \sum \mathbf{c}_q(t) y_1^{m_1} y_2^{m_2} y_3^{m_3} y_4^{m_4} \\ \sum \mathbf{d}_q(t) y_1^{m_1} y_2^{m_2} y_3^{m_3} y_4^{m_4} \end{Bmatrix}, \quad \sum_{i=1}^{4} m_i = 3
$$

where $\mathbf{a}_q(t)$, $\mathbf{b}_q(t)$, $\mathbf{c}_q(t)$ and $\mathbf{d}_q(t)$ once again are vector periodic coefficients with period $2T$. It is observed that the eigenvalue corresponding to the third state is zero and the remaining eigenvalues have negative real parts. The centre manifold relations for this case can be assumed in the form

$$
y_1 = \mathbf{B}_{11}^3(t) y_3^3; \quad y_2 = \mathbf{B}_{21}^3(t) y_3^3; \quad y_4 = \mathbf{B}_{31}^3(t) y_3^3,
$$

(63)

where $\mathbf{B}_{11}^3(t)$, $\mathbf{B}_{21}^3(t)$ and $\mathbf{B}_{31}^3(t)$ are unknown coefficients with period $2T$. These can be determined by solving the differential equations in $\mathbf{B}_{11}^3(t)$, $\mathbf{B}_{21}^3(t)$ and $\mathbf{B}_{31}^3(t)$ as described
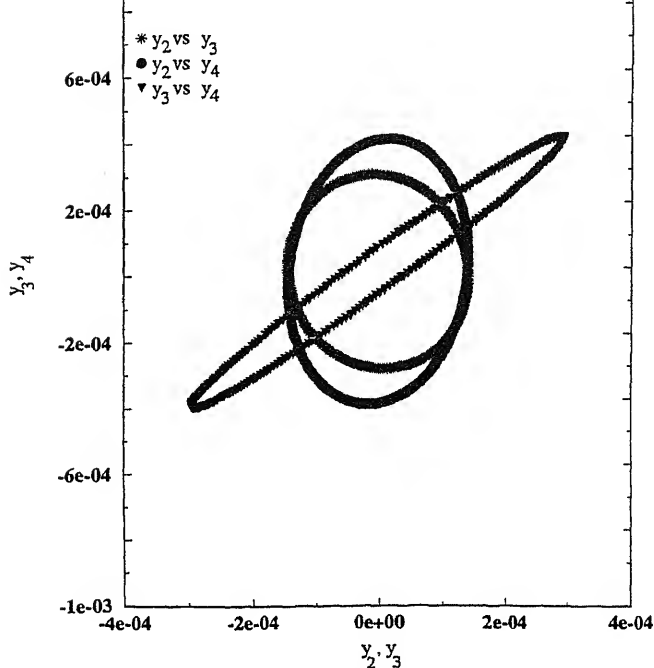
Figure 5. (b) Poincaré plots – Double inverted pendulum under Hopf bifurcation.

earlier. The computed coefficients are recorded by Pandiyan (1994). Substituting (63) into equation (62) and neglecting the higher order terms beyond the cubics, the one-dimensional centre manifold equation is found and since the mean value of the periodic coefficient of the reduced centre manifold equation for this case is positive, the fixed point is unstable and hence the corresponding $2T$ periodic orbits in the original coordinates are unstable. In a similar way, the fold bifurcation of this system can also be studied. However, the results are not included here.

References

Arnold V I 1988 *Geometrical methods in the theory of ordinary differential equations* (New York: Springer-Verlag)

Awreicewicz J 1989 *Bifurcation and chaos in simple dynamical systems* (Singapore: World Scientific)

Bernussou J 1977 *Point mapping stability* (New York: Pergamon)

tions (New York: Gordon and Breach)

Bramwell A R S 1976 *Helicopter dynamics* (London: Edward Arnold)

Bruno A D 1989 *Local methods in nonlinear differential equations* (Berlin, Heidelberg: Providence: Springer-Verlag)

Carr J 1981 *Applications of center manifold theory* (New York: Springer-Verlag)

Chow S N, Wang D 1985 Normal forms of bifurcating periodic orbits. *Contemporary Mathematics. Vol. 56* (Proceedings of a summer research conference, July 1985) (eds) M Golubitsky, J Guckenheimer (Providence, RI: Am. Math. Soc.)

Coddington E A, Levinson N 1955 *Theory of ordinary differential equations* (New York: McGraw Hill)

Cullum J, Willoughby R A (eds) 1986 A practical procedure for computing eigenvalues of large sparse nonsymmetric matrices. *Large scale eigenvalue problem* (New York: Elsevier Science)

Doedel E J, Kernévez J P 1986 AUTO: Software for continuation and bifurcation problems in ordinary differential equations. Applied Mathematics Report, California Institute of Technology

Flashner H, Hsu C S 1983 A study of nonlinear periodic systems via the point mapping method. *Int. J. Numer. Meth. Eng.* 19: 185–215

Floquét G 1883 Sur les équations differentials linéaires a coefficients périodiques. *Annales Scientifiques de l'Ecole Normale Superieue* series 2, vol. 12

Friedmann P, Hammond C E, Woo T H 1977 Efficient numerical treatment of periodic systems with application to stability problems. *Int. J. Num. Meth. Eng.* 11: 1117–1136

Gaonkar G H, Peters D A 1986 Review of Floquet theory in stability and response analysis of dynamic systems with periodic coefficients. *Recent trend in aeroelasticity, structures, and structural dynamics* (ed) P Hajela (Gainsville, FL: University of Florida Press) p 101

Gaonkar G H, Simha Prasad D S, Sastry D 1981 On computing Floquet transition matrices of rotorcraft. *J. Am. Helicopter Soc.* 26: 56–61

Guttalu R S, Flashner H 1989 Periodic solutions of nonlinear autonomous systems by approximate point mappings. *J. Sound Vibr.* 129: 291–311

Guttalu R S, Flashner H 1990 Analysis of dynamical systems by truncated point mapping and cell mapping. *Nonlinear dynamics in engineering systems* (ed) W Schiehlen (New York: Springer-Verlag)

Guckenheimer J, Holmes P 1983 *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields* (New York: Springer-Verlag)

Hale J, Kocak H 1991 *Dynamics and bifurcations* (New York: Springer-Verlag)

Hsu C S 1987 *Cell-to-cell Mapping* (New York: Springer-Verlag)

Hsu C S, Cheng W H 1973 Applications of the theory of impulsive parametric excitation and new treatments of general parametric excitation problems. *J. Appl. Mech.* 40: 78–86

Hsu C S, Cheng W H 1974 Steady-state response of a dynamical system under combined parametric and forcing excitations. *J. Appl. Mech.* 41: 371–378

Jin J-D, Matsuzaki Y 1988 Bifurcations in a two-degree-of-freedom elastic system with follower forces. *J. Sound and Vibr.* 126: 265–277

Johnson W 1980 *Helicopter theory* (Princeton, NJ: University Press)

Joseph P, Pandiyan R, Sinha S C 1993 Optimal control of mechanical systems subjected to periodic loading via Chebyshev polynomials. *Optimal control applications and methods* pp 75–90

Kretz M 1976 Research in multicyclic and active control of rotary wings. *Vertica* 1: 95–105

Lalanne M, Ferraris G 1990 *Rotordynamics prediction in engineering* (Chichester: John Wiley and Sons)

Lindh K G, Likins P W 1970 Infinite determinant methods for stability analysis of periodic-coefficient differential equations. *AIAA J.* 8: 680–686

Lindtner E, Steindl A, Troger H 1990 Generic one-parameter bifurcations in the motion of a simple robot. *Continuation techniques and bifurcation problems* (ISNM 92) (eds) H D Mittelmann, D Roose (Birkhauser); also in *J. Comput. Appl. Math.* (1989) 26: 199–218

Lukes D L 1982 *Differential equations: Classical to controlled* (New York: Academic Press)

Malkin I G 1962 Some basic theorems of the theory of stability of motion in critical cases. *Stability and dynamic systems, Translations, American Mathematical Society, Series 1* (Am. Math. Soc.) 5: 242–290

Mckillip R M Jr 1985 Periodic control of individual-blade-control helicopter rotor. *Vertica* 9: 199–225

Nayfeh A H 1973 *Perturbation methods* (New York: Wiley)

Pandiyan R 1994 *Analysis and control of nonlinear dynamic systems with periodically varying parameters*. Ph D dissertation, Dept. Mechanical Engineering, Auburn University

Pandiyan R, Sinha S C 1995 Analysis of time-periodic nonlinear dynamical systems undergoing bifurcations. *Nonlinear Dynamics* 8: 21–43

Pandiyan R, Bibb J S, Sinha S C 1993 Liapunov–Floquet transformation: Computation and application to periodic systems. In *Dynamics and vibration of time-varying systems and structures 14th ASME Biennial Conference on Mechanical Vibration and Noise* (eds) S C Sinha, R M Evan-Iwanowski (New York: ASME Press) pp 337–348; also in *J. Vibr. Acoust.* 118: 209–219

Peters D A, Hohenemser K H 1971 Application of the Floquet transition matrix to problems of lifting rotor stability. *J. Am. Helicopter Soc.* 16: 25–33

Poincaré H 1899 *Les Methodes Nouvelles de la Mécanique Céleste* (Paris: Gauthier-Villars)

Rosenblat S, Cohen D S 1980 Periodically perturbed bifurcation I. Simple bifurcation. *Studies in Appl. Math.* 63: 1–23

Rosenblat S, Cohen D S 1981 Periodically perturbed bifurcation II. Hopf Bifurcation. *Studies in Appl. Math.* 65: 95–112

Sanders J A, Verhulst F 1985 *Averaging methods in nonlinear dynamical systems* (New York: Springer-Verlag)

Sethna P R, Schapiro S M 1977 Nonlinear behavior of flutter unstable dynamical systems with gyroscopic and circulatory forces. *J. Appl. Mech.* 44: 755–762

Seydel R 1981 Numerical computation of periodic orbits that bifurcate from stationary solutions of ordinary differential equation. *Appl. Math. Comput.* 9: 257–271

Seydel R 1987 New methods for calculating stability of periodic solutions. *Comput. Math. Appl.* 14: 505–510

Seydel R 1988 *From equilibrium to chaos: Practical bifurcation and stability analysis* (New York: Elsevier)

Sinha S C, Joseph P 1994 Control of general dynamic systems with periodically varying parameters via Liapunov–Floquet Transformation. *ASME J. Dynamic Syst., Measurements Control* 116: 650–658

Sinha S C, Juneja V 1991 An approximate analytical solution for systems with periodic coefficients via symbolic computation. *AIAA/ASME/ASCE/AHS/ASC 32nd Structures, Structural Dynamics and Materials Conference* (A Collection of Papers, Part 1) pp 790–797

Sinha S C, Pandiyan R 1994 Analysis of quasilinear dynamical systems with periodic coefficients via Liapunov–Floquet transformation. *Int. J. Non-Linear Mech.* 29: 687–702

Sinha S C, Wu D-H 1991 An efficient computational scheme for the analysis of periodic systems. *J. Sound Vibr.* 15: 345–375

Sinha S C, Wu D-H, Juneja V, Joseph P 1993a Analysis of dynamic systems with periodically varying parameters via Chebyshev polynomials. *ASME J. Vibr. Acoust.* 115: 96–102

Sinha S C, Senthilnathan N R, Pandiyan R 1993b A new numerical technique for the analysis of parametrically excited nonlinear systems. *Nonlinear Dynamics* 4: 483–498

Stoker J J 1950 *Nonlinear vibration* (New York: Interscience)

Wu D-H, Sinha S C 1994 A new approach in the analysis of linear systems with periodic coefficients for applications in rotorcraft dynamics. *Aeronaut. J. R. Aeronaut. Soc.* January: 9–16

Yakubovitch V A, Starzhinskii V M 1975 *Linear differential equations with periodic coefficients* (New York: Wiley-Halsted) vols. 1 and 2

ectrochemical discharge machining: Principle and ssibilities

AMITABHA GHOSH

Department of Mechanical Engineering, Indian Institute of Technology, Kanpur 208 016, India
Present address: Indian Institute of Technology, Kharagpur 721 302, India
e-mail: amitabha@iitkgp.ernet.in

Abstract. This paper highlights the important results of the investigations on Electrochemical Discharge Machining (ECDM) conducted by the author and his coresearchers. It has been found that "switching phenomenon" plays a crucial role in spark generation and not the straightforward breakdown of the non-conducting vapour blanket. The mechanism of spark generation has been understood reasonably well and inductance in the circuit has emerged as an important process parameter from the investigations. This information has been effectively used to improve the process capability of ECDM by a substantial amount. It has also been shown how ECD can be very conveniently used for micro-welding operation without using any sophisticated arrangement.

Keywords. Electrochemical discharge machining; switching phenomenon; spark generation; inductance; micro-welding.

Introduction

nufacturing is now recognized by all sections of our society as one of the most important vities since it affects very significantly the GNP, international trade and employment. nufacturing is broadly defined as

"an activity by which material, labour (including information), energy and machines are brought together to produce a product whose value is more than the total value of the individual inputs".

s is schematically indicated in figure 1. However, this definition is too generalised and can be more specific stating that
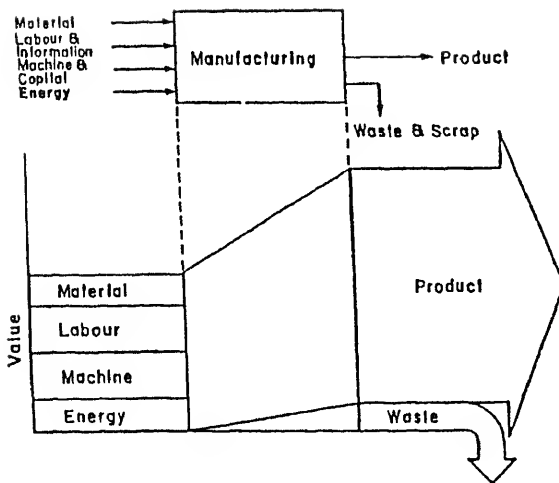
Figure 1. Value addition in manufacturing activity.

"Engineering industry manufacturing" is what we mean by "manufacturing" in the field of Mechanical Engineering. Figure 2 suggests that almost one third of the production in India comes from such activity and, therefore, "mechanical manufacturing" represents a very vital pillar of our economy.

"Machining" has a special status among all processes covering the whole spectrum of "mechanical manufacturing" because of its capabilities as below.

- High accuracy and finish can be achieved.

- Complex shapes can be produced.

- Machining force and power are independent of the work size and shape as all machining processes are basically generating processes.

- The property of the bulk of the work material remains unaffected.

Machining has also been one of the earliest manufacturing processes and is continuously evolving to cope with the new challenges faced by the manufacturing industries.
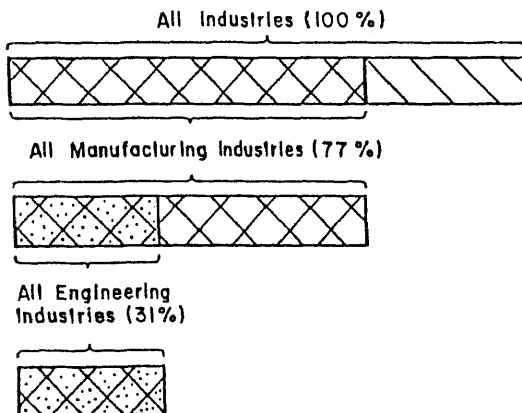
Figure 2. Proportion of engineering industries.

Removal Mechanisms
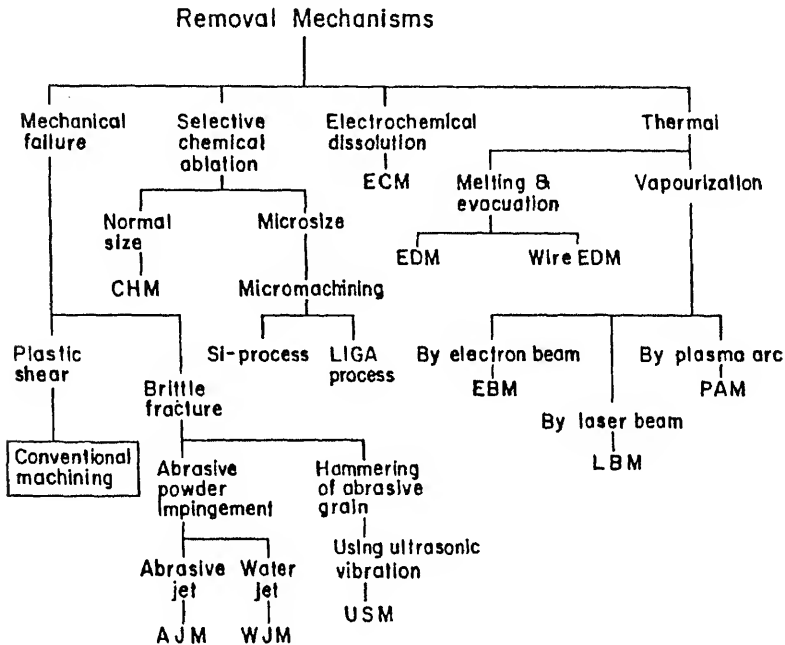
Figure 3. Family of machining processes.

In machining, the required shape, size and finish are obtained by removing excess material in a controlled manner in the form of small chips. In conventional machining removal is in the form of small chips and in many advanced processes chip size is extremely small. Removal of excess material can be achieved in various ways using different forms of energy as shown by the tree diagram in figure 3.

The two most extensively used unconventional machining processes are

- Electrochemical machining (ECM), and

- Electric discharge machining (EDM).

Considering the different aspects of their capabilities, these two processes are considered to possess maximum potential. However, inspite of their many virtues both these processes suffer from a very important limitation. To employ either ECM or EDM the work material must be *electrically conducting*. The development of electrochemical discharge machining has taken place in recent years primarily to eliminate this difficulty.

2. Electrochemical discharge machining

A study of all machining processes indicates that any scheme for removing material in the form of small particles in a controlled fashion can be used for the shaping of objects. To circumvent material and shape problems, quite often the approach adopted for machining is to cause the melting of a small portion of the workpiece by means of intense localized heat generation. By controlling the location of the heat source in a proper way, the required shape of the workpiece can be achieved.
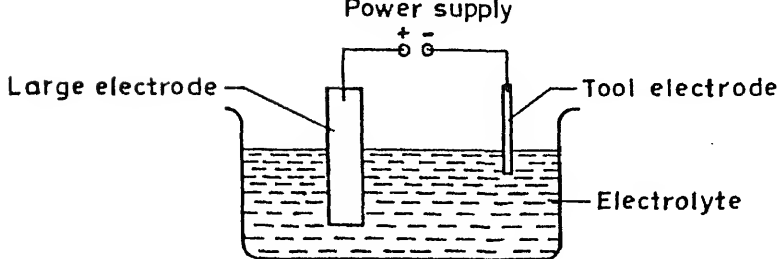
Figure 4. Schematic diagram of ECD set-up.

One phenomenon used to produce intense localized heat generation without using any sophisticated technology like electron beam or laser beam is the electrochemical discharge. Figure 4 shows an electrochemical cell where one electrode is very small and the other is relatively much larger. When these electrodes are connected to a voltage source (either AC or DC) an electrical discharge can be seen at the tip of the smaller electrode if the supply voltage exceeds a critical value. This is known as electrochemical discharge which takes place between the tip of the smaller electrode and the electrolyte in its immediate neighbourhood.

The temperature at the discharge zone has been measured by thermocouple technique (Basak 1991) and the result is shown in figure 5. This is, of course, the temperature of the thermocouple junction which itself happens to be the tip of the smaller electrode. The temperature of the discharge itself has been estimated by emission spectroscopy (Reghuram 1994) for different electrolytes with varying concentration. The temperature is found to be in the range 8000–10 000 K.

Thus, it is possible to cause material removal in very small quantities by melting and evacuation if the workpiece is kept very near the tool tip and within the range of the discharge. The work material *need not be electrically conducting in this case*. Such a scheme is presently in the process of development and this process is termed "Electrochemical Discharge Machining (ECDM)."

The scheme of the basic ECDM process is indicated in figure 6. The tip of the tool electrode is very close to the work surface (electrically nonconducting) and small holes can be drilled by this process. Slots can be also cut by using a knife edge tool electrode. Slots or grooves of any shape can be machined by controlling the lateral motion of the tool tip of a pointed tool.

3. Mechanism of ECD phenomenon and ECD machining

This process, though it appears to have considerable potential for machining electrically non-conducting materials, has very limited acceptance mainly because of its limited capacity. To improve the process capability of ECDM it is essential to understand the basic mechanism of ECDM and identify the process parameters correctly.

The process is extremely complex involving electrochemistry, heat transfer and boiling, melting and evacuation. To understand ECDM, first the phenomenon of ECD should be investigated. For this purpose an extensive study has been carried out. One of the important

(a)

(b)
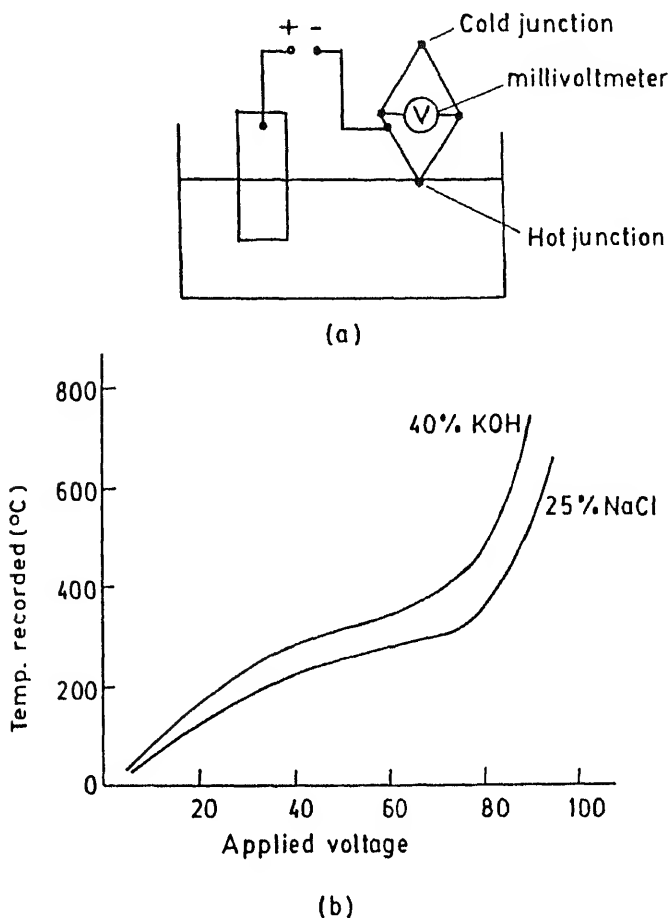
Figure 5. (a) Set-up to determine electrode temperature. (b) Result obtained.

findings is the nature of the potential drop across the electrolytic cell. A typical variation is indicted in figure 7 (Allesu 1988). It is clear that an overwhelmingly large fraction of the total drop takes place across a very thin layer surrounding the tool electrode (the smaller one) during electrochemical discharge. This suggests that a resistant layer develops around the tool electrode. In most experiments this is the cathode and, therefore, H_2 evolves in the form of very fine bubbles. Moreover, the temperature of the electrolyte at this region increases because of much higher current density. This leads to boiling if the current density is high enough. Both these result in the development of a non-conducting gas and vapour blanket around the smaller electrode (this does not happen at the larger electrode because the current density is much less). When the voltage (either full-wave rectified DC or smooth DC) applied across the ECD cell is gradually increased, the corresponding change in the average current is shown in figure 8. An extensive study shows (Allesu 1988; Basak 1991; Allesu *et al* 1992; Reghuram 1994) that these characteristics remain unchanged for all situations. It is seen that electrochemical discharge starts only when the applied voltage reaches a critical value which depends on the type of electrolyte and its concentration. The
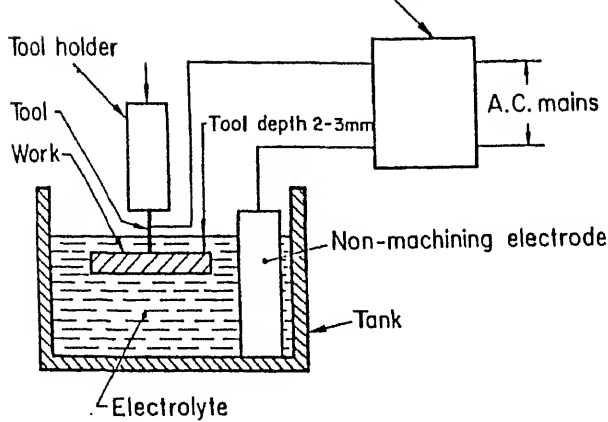
Figure 6. Configuration of non-conductor work machining using ECD.

depth to which the electrode is submerged and the electrode diameter do not appear to have much effect as long as the discharging electrode is much smaller than the other one. It is also very interesting to note that the power consumption (for a given electrolyte and set of electrodes) at the critical condition ($= V_{cr}.I_{cr}$) remains constant and is independent of electrolyte concentration (Basak 1991). This strongly suggests that blanketing can be a major factor in the occurrence of electrochemical discharge as both H_2 evolution and vapour production through boiling depend on the power only.

The conditions under which blanketing takes place have been investigated and a theoretical model has been developed for the ECD mechanism and for ECDM (Basak 1991; Basak & Ghosh 1995) which can predict the critical voltage, critical average current and also the material removal rate under prescribed conditions. The results agree reasonably well with the experimental observations. This investigation also provides a new input which suggests that the electric discharge is primarily due to (Basak 1991; Basak & Ghosh 1995) switching action and not to the breakdown of gas in the blanket as suggested by the previous researchers.

Figure 9 shows the various stages of bubble coverage of the tool surface as the applied voltage is gradually increased. At the critical condition, the hemispherical bubbles cover the surface completely. In this closed-packed condition, the in-between conducting bridges are very narrow and the current densities in these bridges become very high and cause instant boiling which blows the bridges. When a bridge is blown a spark is generated by switching action. This is why sparking takes place even if the applied voltage is much lower than the minimum voltage for breakdown as indicated in the Paschen curve. Figure 10 shows the equivalent circuits. As switching action plays the crucial role in spark generation, inductance in the circuit becomes one of the most important parameters. This however remained unnoticed so far. This finding thus leads to an extra process parameter which can be easily controlled and a substantial improvement in the process capability of ECDM can be achieved.
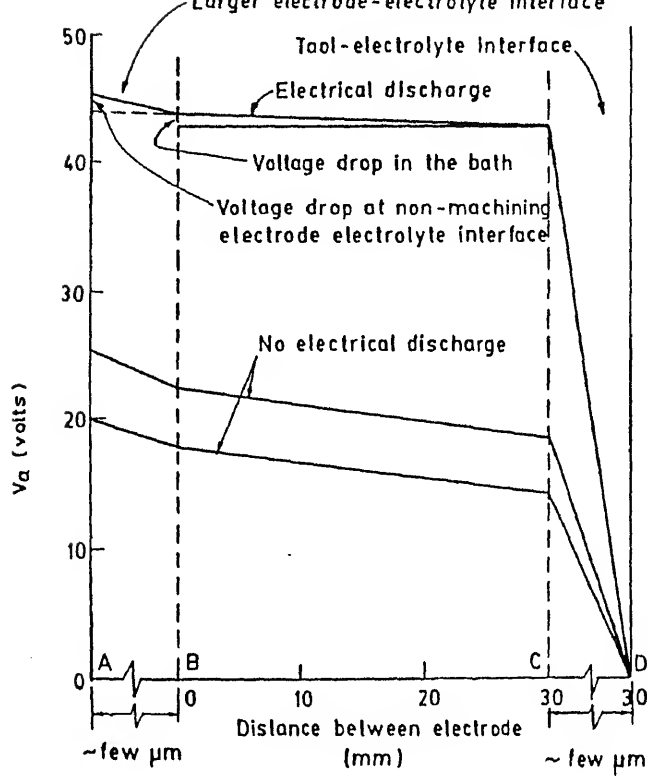
Figure 7. Distribution of voltage drop in an ECDM bath [Electrolyte : NaOH (35%); tool polarity: −ve; power supply : DC].
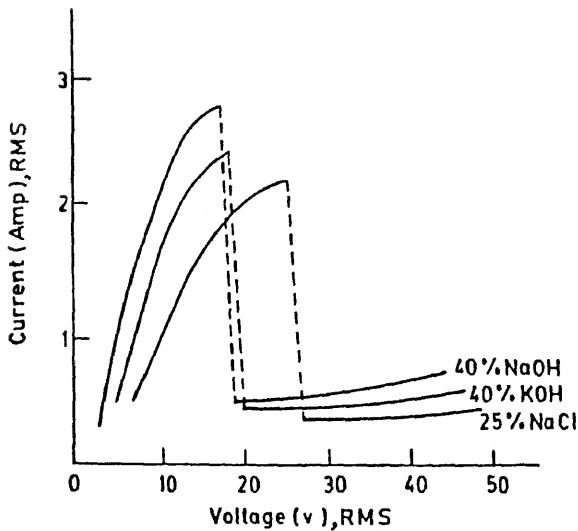
Tool dia: 0·11 cm   Tool depth :=0·2 cm
Power supply : Smooth D.C.

Figure 8. V-I characteristics for different electrolytes.
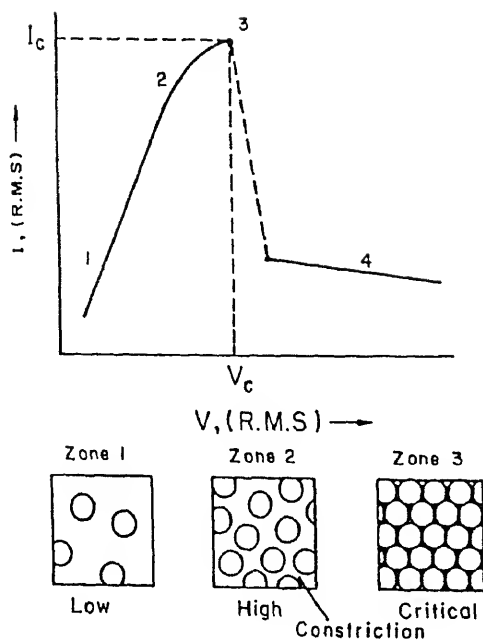
Figure 9. Bubble density on tool electrode at different applied voltages.
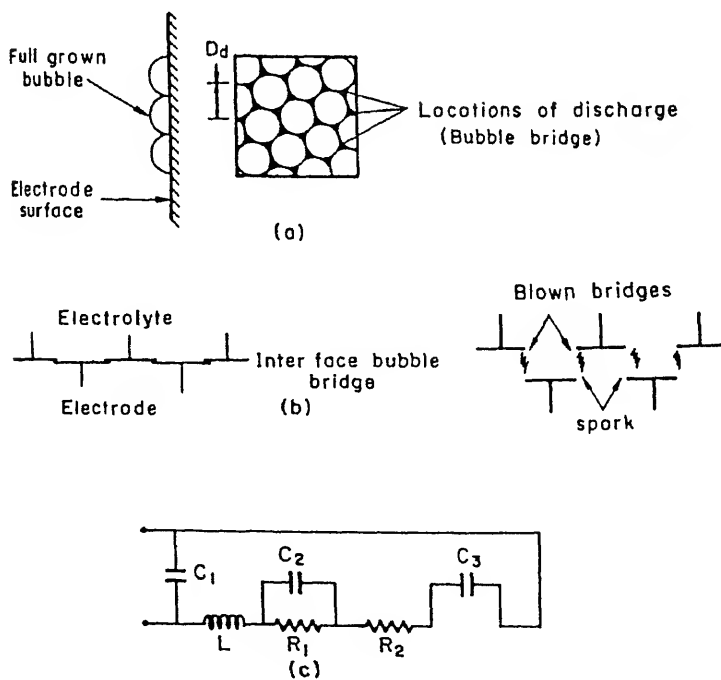
Figure 10. (a) Discharge locations with bubble distribution at critical condition. (b) Idealised switching off situation, and (c) Idealised equivalent circuit at discharge.
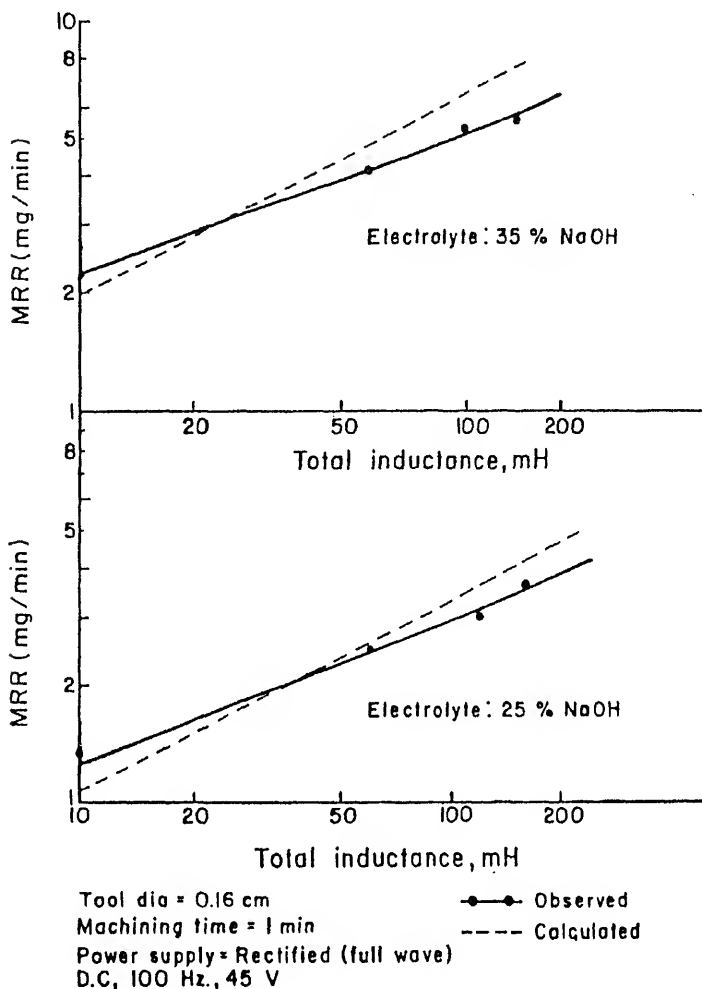
Figure 11. Total inductance vs MRR (NaOH) (log scale, workpiece – glass).

4. Improvement of process capability

A number of experiments on ECDM have been conducted using artificially introduced inductance in the circuit (Basak 1991; Basak & Ghosh 1995b). It has been observed that increasing the inductance in the circuit results in higher material removal rate. The increase is quite significant and a set of sample results is shown in figure 11. Further investigation can provide more interesting results in this direction.

The idea of using artificially introduced inductance in an ECDM circuit has been put forward (Reghuram 1994) in a somewhat different direction. It has been found that when smooth DC voltage is applied, the introduction of a series inductance of adequate value in the circuit can lead to a "total discharge condition." In this condition the discharge appears like a flame emerging from the tool tip. To obtain a stable total discharge without the undesirable disturbance in the electrolyte, the input voltage is increased to a high value (≈ 100 V) to obtain stable discharge. Once the stable discharge is attained the input voltage
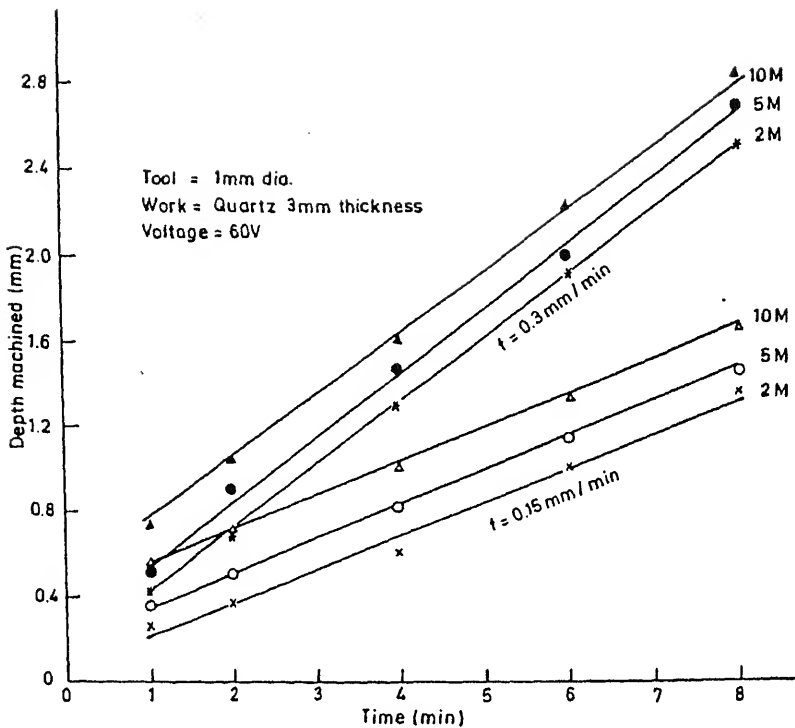
Figure 12. Variation of machined depth with time [electrolyte: KOH].

is lowered to the operational value (≈ 60 V) and it is found that the energy input to the system is much higher.

Under this situation it is not necessary to keep the workpiece in contact with the tool tip as is done in the common ECDM. It is enough to keep the work surface within the range of the stable discharge; the process can be considered to be somewhat analogous to laser-beam machining. However, gravity feed is no longer possible in this type of ECDM and the tool has to be fed by a servo drive as is done in EDM. Figure 12 shows that by employing this modified ECDM holes can be drilled in a quartz block with a penetration rate of 0.3 mm/min and can be continued even if the tip reaches depth of the order of 3 mm or more. This has not been possible with the usual ECDM without a series inductance.

5. Other manufacturing applications

At this point it is worth mentioning a few other applications of ECD in manufacturing processes. Using this phenomenon a scheme for scribing on non-conducting materials has been developed (Allesu 1988). A scribing pen to write on glass and ceramics was fabricated and it performed satisfactorily.

Another important manufacturing application of ECD phenomenon besides machining is microwelding. It has been shown (Allesu 1988) that micro-welding of a fine thermocouple can be very conveniently performed using the heat generated by ECD. The scheme is
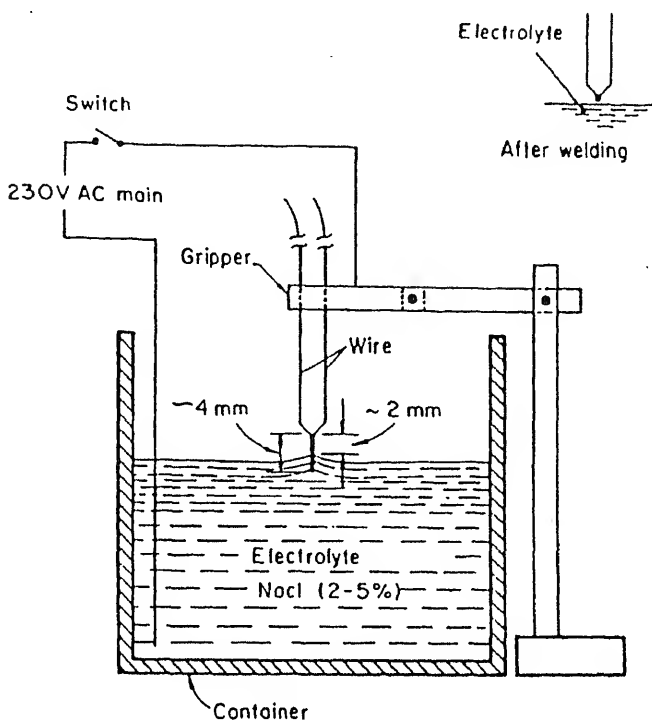
Figure 13. Schematics of micro-welding set-up.

very simple and is shown in figure 13. The process has been investigated in detail (Parija 1993; Ghosh *et al* 1995) and it was possible to determine the optimal combinations of the process parameters for obtaining excellent welds. Figure 14 shows the suitable regions of the voltage-current combination which yield good results. The work is still in progress.

Since a controlled localized heating is possible with the help of electrochemical discharge other important processing methods can be developed. Surface treatment using the pointed heat source is one such example. A more interesting application of the ECD phenomenon can be "Rapid Prototyping" with fused deposition of very thin metallic wires as indicated schematically in figure 15. Currently rapid prototyping using metals instead of polymers, resins, plastics etc. are done using high power lasers and the system is expensive. A low budget metallic fused deposition modelling type rapid prototyping may be possible using ECD.

6. Concluding remarks

The present investigations indicate that ECD can be a very useful tool for different types of manufacturing processes. The mechanism of spark generation is found to be primarily the switching phenomenon rather than the breakdown of insultaing gas layer. This understanding resulted in a major improvement of the process capability by artificially introducing a suitable amount of inductance in the circuit. Further research resulted in a
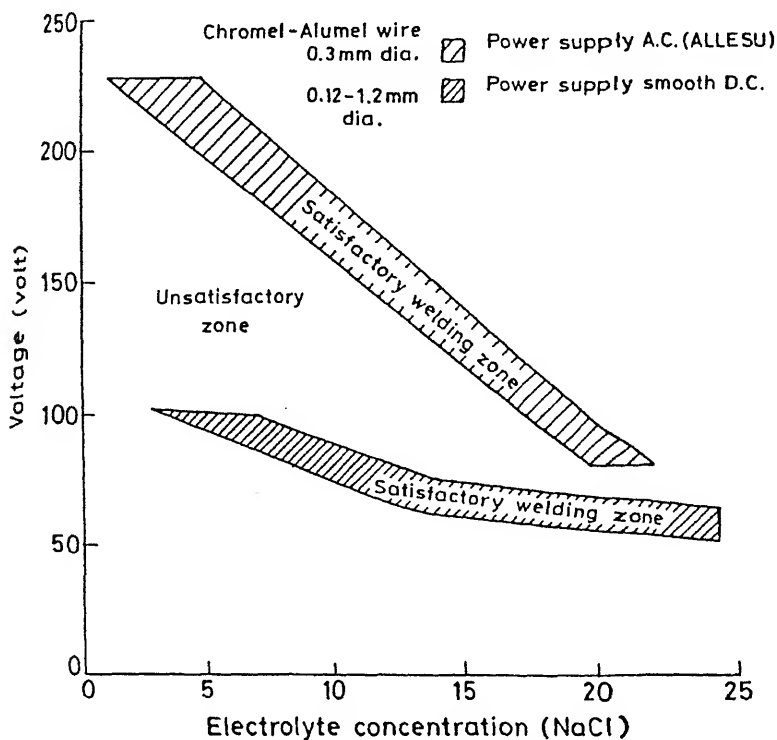
Figure 14. Nature of voltage – electrolyte concentration in microwelding.
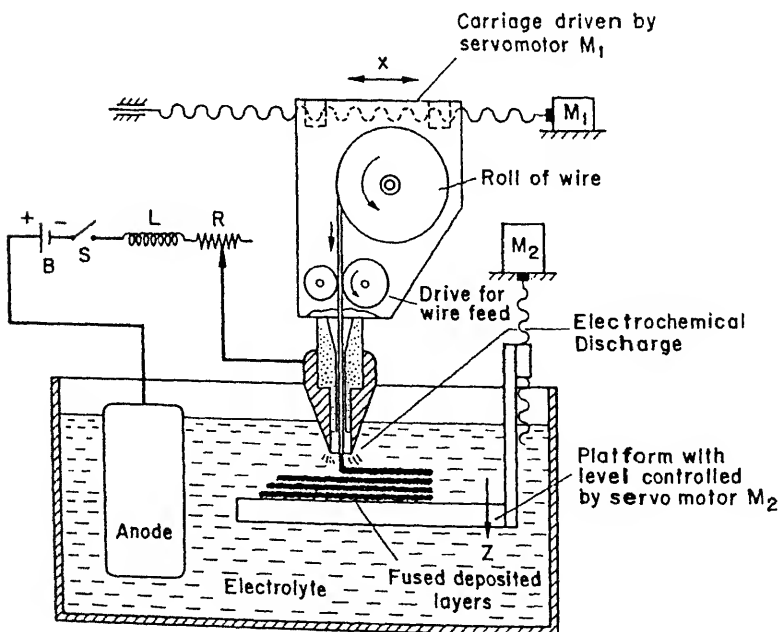
Servodrive for the y– direction not shown

Figure 15. Scheme of ECD fused deposition modelling.

D phenomenon has been very effectively used for micro-welding. ECD micro-welding very simple process and can be useful to the industry. Similarly there is a great poten- for ECD in other types of material processes and fused deposition modelling for rapid totyping.

ferences

esu K 1988 *Electrochemical discharge phenomenon in manufacturing processes*. Ph D thesis, Indian Institute of Technology, Kanpur

esu K, Ghosh A, Muju M K 1992 A preliminary quantitative approach of a proposed mechanism of material removal in electrical machining of glass. *Eur. J. Mech. Eng.* 36: 202–207

ak I 1991 *Electrochemical discharge machining mechanism and a scheme for enhancing material removal capacity*. Ph D thesis, Indian Institute of Technology, Kanpur

ak I, Ghosh A 1995a Mechanism of spark generation during electrochemical discharge machining. *J. Mater. Process. Technol.* 62: 46–53

ak I, Ghosh A 1995b Mechanism of material removal in electrochemical discharge machining: A theoretical model and experimental verification *J. Mater. Process. Technol.* (in press)

osh A, Muju M K, Parija S, Allesu K 1995 Microwelding using electrochemical discharge. *Int. J. Machine Tools Manuf.* (in press)

ija S 1993 A thermal model and parametric investigation of electrochemical discharge microwelding. M Tech thesis, Indian Institute of Technology, Kanpur

ghuram V 1994 *Electrical and spectroscopic investigations in electrochemical discharge machining*. Ph D thesis, Indian Institute of Technology, Madras and Indian Institute of Technology, Kanpur

chining and surface integrity of fibre-reinforced plastic
nposites

M RAMULU

Department of Mechanical Engineering, University of Washington, Seattle,
WA 98195, USA
e-mail: ramulum@me.washington.edu

Abstract. The current focus of manufacturing research on fibre-reinforced
plastics (FRP) is composed of the search for efficient processing techniques
capable of providing high quality machined surfaces. Very limited work has
been performed to identify the influence of manufacturing processes like edge-
trimming and drilling on material performance. Recent reports suggest that
process-induced damage may affect the mechanical behaviour of FRP materi-
als. Therefore an experimental study of orthogonal cutting was conducted on
the edge trimming of unidirectional and multi-directional graphite/epoxy com-
posites with polycrystalline diamond tools. The effects of tool geometry and
operating conditions were evaluated from an analysis of chip formation, cutting
force, and machined surface topography. All aspects of material removal were
found to be primarily dependent on fibre orientation. Discontinuous chip for-
mation was noted throughout this study, regardless of machining parameters.
Three distinct mechanisms in the edge trimming of fibre-reinforced composite
material including a combination of cutting, shearing, and fracture along the
fibre/matrix interface were observed. An investigation conducted on the com-
pression, flexural and impact strength of graphite/epoxy composites machined
by both traditional and non-traditional techniques, confirms that manufactur-
ing characteristics may not only affect bulk properties but also influence the
initiation and propagation of failure.

Keywords. Fibre-reinforced plastics; manufacturing processes; process-
induced damage; graphite/epoxy composites.

Introduction

e-reinforced plastics (FRP) are a contemporary class of multi-constituent materials
h exhibit superior strength and stiffness-to-weight ratios in comparison to homoge-
s counterparts. Consequently, polymeric composites are quickly gaining prominence
e aerospace, automotive, and other high technology industries, where both weight and

performance govern design constraints. The ability to develop complicated part shapes and align fibres to accommodate for critical load orientations has facilitated the design and production of high-performance aircraft using composite materials in excess of 50 percent (Kinkaid 1988). Construction of commercial and military aircraft composed entirely of composite materials is imminent (Riggs 1984; McCarty 1993). However, specific details concerning post-mold manufacturing technology utilized for FRPs inhibit their use in new designs and in the development of new products. Contrary to net-shape manufacturing of homogeneous materials, FRPs are traditionally molded to near net shape with complicated patterns incorporating specific geometry and laminae orientation, then cured with a combination of temperature and pressure. Secondary processing such as trimming to final shape and drilling for fasteners is typically required to facilitate component assembly.

Surface roughness and tolerance are closely related, and it is generally necessary to specify a smooth finish to maintain a fine tolerance in the finishing process. For many practical design applications, tolerance and strength requirements impose a limit on the maximum allowable roughness. The reliability of machined components, especially for fibre-reinforced plastics (FRP) in high strength applications, is often critically dependent upon the quality of surface produced by machining, and the surface layer may drastically affect the strength and chemical resistance of the material. However, inhomogeneity of FRPs, caused by the difference in the properties of the fibre and the matrix materials, will have a machined surface that is less regular and is usually rougher in comparison with machined metal surfaces (Koplev et al 1983; König et al 1985; Abrate & Walton 1992a).

Traditional methods of machining often induce critical flaws in the component parts during net trimming, and various degrees of delamination, splintering, fibre pullout, and cracking have been reported (Koplev et al 1983; König et al 1985; Ho-Cheng & Dharan 1990; Abrate & Walton 1992a; Colligan & Ramulu 1992; Wang et al 1992). Due to the limitations of traditional processes, many non-traditional manufacturing methods have been examined for net shape machining of composites (Koplev et al 1983; König et al 1985; Ho-Cheng & Dharan 1990; Abrate & Walton 1992a; Colligan & Ramulu 1992; Wang et al 1992). The abrasive waterjet (AWJ) is one such method which appears to be highly suited for production trimming of FRP materials (Ramulu & Arola 1993, 1994). Owing to its ability to achieve a quality surface at rapid production rates, AWJ is currently being sought for production applications (Mortimer 1987; Miller 1991). However, very little is known regarding the effects of AWJ machining on the surface integrity of FRPs, or the influence of post-mold manufacturing defects on mechanical performance in general. Considering the present and future applications of FRP materials to commercial aerostructures and other related industries, it is necessary to obtain a fundamental understanding of post-mold manufacturing process effects on the structural integrity of fibre-reinforced materials. Investigating by machining-incurred damage to mechanical performance and service life is imperative to the reliability of composite structures.

The choice of net trimming or drilling technique is often based on a combination of process economics and visual aesthetics of the machined surface. Although important, these factors are not appropriate for the choice of a manufacturing method considering the probable influence of secondary operations on the performance of composite materials. Surface roughness characteristics and surface integrity of machined FRP composites in general, and graphite/epoxy in particular, have not been studied in detail. Therefore, recent

s concerning machining-induced edge effects on the structural performance of FRPs
viewed and results from an investigation on machining mechanisms and machining
ed surface integrity of graphite/epoxy will be presented.

ackground

Machined quality of fibre-reinforced plastics

factured surfaces are evaluated using a number of different methods, each of which
nique quantitative features corresponding to the machined quality. The choice of
ction technique may depend on a variety of criteria including available equipment,
vestigator, and other terms in which quality control guidelines are established. Me-
cal performance of homogeneous material component parts has been shown to be
ction of the residual stress induced in the material and the surface topography.
reinforced plastics do not develop residual stress by applied forces. Therefore, the
y of machined FRP parts is often interpreted by surface profilometry and visual
ques.

Surface morphology: Many methods have been used in evaluating the quality of
rimmed surfaces such as the use of average roughness parameters, c-scans, and other
 techniques. Some of these methods and associated quality criteria may be insuffi-
for distinguishing features that influence part performance (Philips & Parker 1987;
old *et al* 1989; Jamil & Chambers 1991; Wern 1991; Ramulu *et al* 1993). For in-
, visual inspections are often used to identify the degree of damage to a component
nd also to discuss prevalent topographical features. However, visual documentation
les little information that can be used in a quantitative analysis or non-subjective
arison. Ramulu and Wern (Wern 1991; Ramulu *et al* 1993) studied the suitability of
rd roughness parameters in describing the topography of unidirectional and multi-
ional graphite/epoxy (Gr/Ep) trimmed with polycrystalline diamond (PCD) tools.
us statistical and random process methods were used in addition to standard rough-
parameters to analyse the trimmed surface characteristics. In this study it was found
ithmetic average surface roughness (R_a), which is often used as a parameter for qual-
ntrol, failed to describe the relative extent of surface variation observed from visual
sis. More appropriate parameters for describing topographical features of the inho-
neous material surface include the peak-to-valley height (R_y) and ten-point height
Cumulative height distribution (CHD) on a probit scale and power spectrum density
) function were also found very useful in characterizing the spatial distribution of the
rimmed graphite/epoxy material. However, one drawback of a profile based descrip-
 that matrix smearing resulting from traditional edge trimming obliterates details of
mmed surface (Wang 1993). This implies that even in rigorous quality inspection of
raphy, some features remain unnoticed that may influence the structural integrity of
ponent part. Therefore, it is often desirable to supplement quantitative descriptions
ded by profilometry with a visual analysis of the surface morphology.

2.1b   *Microstructural integrity*:   Post-processing microstructural integrity refers to the relationship between fibre and matrix after machining. Research addressing microstructural integrity resulting from post-mold processing of FRP's is limited. Preliminary work has shown that AWJ machining of Gr/Ep does not affect the interfacial relationship between constituents whereas studies on traditional trimming and laser cutting of FRP's suggests that the fibre/matrix interface may be altered (Tagliaferri *et al* 1985; Wern 1991; Ramulu & Arola 1993). Microstructure of the machined surface and accumulation of process-induced damage could prove crucial to the structural integrity of FRP's, especially when subjected to cyclic loading. In an assessment of post-mold processing influence of composite materials, microstructure should be examined and correlated with mechanical strength.

2.2   *Machining-induced edge effects on structural integrity*

The design of a FRP component is typically based on choice of constituents, distribution and orientation of the laminae, and the geometry chosen to accommodate for service loads. In addition to these parameters, the effects of manufacturing and associated defects must be considered.

2.2a   *Drilling effects*:   A handful of studies have been reported on the influence of drilled hole quality to the structural integrity of FRP materials (Wood 1978; Pengra & Wood 1980; You & Chou 1988; Ghasemi Nejjad & Chou 1990; Tagliaferri *et al* 1990; Lin & Lee 1992; Mehta *et al* 1992). Some alternatives to conventional drilling like molding holes prior to the curing process have been suggested to improve part strength of components requiring fastener holes. For instance, in a comparison with molded holes, it was recently shown that laminates with drilled holes exhibit lower tensile, compressive, and bearing strengths (You & Chou 1988; Ghasemi Nejjad & Chou 1990; Lin & Lee 1992), with a reduction ranging between 20 and 70%. The superior strength of panels with molded holes has been attributed to fibre continuity, an increase in fibre volume fraction near the hole vicinity, and the absence of matrix microcracks induced during drilling (Ghasemi Nejjad & Chou 1990).

There are remarkable discrepancies between results from different investigations reported on the influence of drilling-induced hole quality to static and fatigue strength of FRP panels (Wood 1978; Pengra & Wood 1980; Tagliaferri *et al* 1990; Mehta *et al* 1992). Wood (1978) concluded that hole quality does in fact influence the fatigue strength of composite panels but does not noticeably influence the tensile strength. Pengra & Wood (1980) similarly concluded from experimental work that restrictions on chipout defects (pitting) and oversize holes could not be relaxed (from definitions in 1980) but suggested that quality control requirements for exit delamination in graphite/epoxy may be too stringent. On the other hand, Mehta *et al* (1992) found that hole quality in PMR-15/Gr laminate, particularly including exit ply delamination, may influence both tensile and compressive properties of panels with drilled holes. Tagliaferri *et al* (1990) concluded that although hole quality does not influence the tensile properties of composites, it has significant influence on bearing strength. Although results from these studies are inconclusive, they unquestionably suggest that drilling-induced fastener hole quality influences the mechanical behaviour of FRP materials. Since drilled-hole anomalies are consistent with those generated during traditional

ge-trimming of FRP's, it should be rationalized that net-trimming operations and man-
acturing induced defects may influence the structural integrity of FRP component parts.

2b   *Edge finishing effects*:   In addition to the choice of constituents, one of most im-
rtant parameters affecting the mechanical properties of a FRP laminate material is com-
nent lay-up. Pipes and Pagano (1970) studied the influence of stacking sequence on
e development of interlaminar shear and interlaminar normal stresses in FRP laminates
bjected to external loading. These phenomena are commonly referred to as "free edge
ects". Edge effects coupled with the presence of manufacturing defects on the trimmed
ge may prove critical to component performance. Large magnitude interlaminar stresses
the free edge necessary to compensate for force and moment equilibrium between plies
creases the need for maintaining high edge quality during net-shape trimming. Edge trim-
ng damage of FRP component panels could be critical to the stress transfer mechanisms
the free edge.

Howarth and Strong (1990) investigated the influence of machining induced edge effects
abrasive waterjet (AWJ) and laser cutting on the strength of FRP materials. Tensile spec-
ens of various widths were machined with both laser and AWJ methods from fiberglass,
evlar, and graphite fibre epoxy matrix composites. In tests conducted with Kevlar fibre
mposites, tensile strength and modulus decreased with decreasing specimen width as-
ciated with the influence of manufacturing induced edge effects. However, from testing
fiberglass composites, specimens machined with the laser decreased with decreasing
ecimen width but abrasive water jets specimen tensile strength increased with decreas-
g specimen width. Graphite fibre composites machined with the AWJ showed the same
nd, increasing tensile strength with decreasing specimen width. Laser machined Gr/Ep
ecimens were not tested due the poor cut quality resulting from thermal properties of the
aphite fibres. No explanation was provided to support the unusual experimental results
d no subsequent studies have been reported. Sadat (1988), on the other hand, examined
e influence of cutting parameters on process damage and resulting interlaminar tensile
ength of unidirectional Gr/Ep composites machined with a 60 tooth high speed steel
SS) slot saw. From an optical study of the machined surface, damage features in the
m of cracks, delamination, and fibre rotation were noted within the machined region.
e extent of machining damage decreased with increasing cutting speed corresponding
the decrease in cutting forces monitored in both the feed and normal directions. Fur-
ermore, minimizing machining damage with high cutting speeds resulted in an increase
the interlaminar tensile strength.

Based on this brief review of machining effects on the surface integrity of FRP materials,
appears that the choice of processing techniques as well as the process conditions used
machining may be significant. However, only a limited degree of work has been done in
s area, and current knowledge of machining influence on surface integrity is insufficient.
ditional investigations are necessary to enhance the understanding of manufacturing
ects on FRP material performance. Therefore, an investigation on the surface integrity
graphite/epoxy was conducted at the University of Washington was started about ten
ars ago to study the influence of edge-trimming with polycrystalline diamond tools,
amond abrasive cutters, and abrasive waterjets on the compression and flexural strength
graphite/epoxy composites.
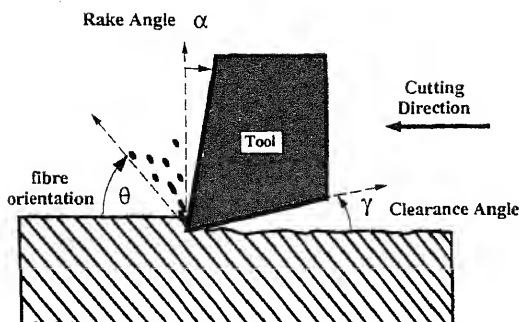
Figure 1. Fibre orientation and insert geometry.

3. Orthogonal machining

3.1 *Materials and procedures*

A unidirectional 4mm thick graphite/epoxy (Gr/Ep) panel with 3501-6 resin and IM-6 fibres was used in this investigation. A 1 m square unidirectional panel was fabricated into test specimens with the desired fibre orientations for orthogonal trimming. Fibre orientations are defined clockwise with reference to the cutting direction as shown in figure 1. Unidirectional fibre orientations greater than 90° are typically regarded as negative orientations by convention. A multidirectional Gr/Ep laminate panel with 6 μm fibre diameter, 0.68 volume fraction, and 200 μm ply thickness was used for the edge-trimming experiments. Panel layup of $[45°/-45°/(0°/90°/45°/-45°)_2]_s$ eliminated stretching/bending coupling and resulted in less chance of delamination and debonding. PCD tool inserts in the trimming experiments were medium grade General Electric COMPAX 1300 with 7 μm grain size. Various tool insert geometries were studied including 0°, 5°, and 10° rake angles (α) with 7°, and 17° clearance angles (γ). Mechanical properties of the PCD material and unidirectional Gr/Ep are listed in table 1.

A Rockford planer-shaper equipped with a hydraulic table to provide steady cutting motion was used for the orthogonal edge-trimming experiments. On-line dynamic cutting

Table 1. Mechanical properties of PCD and Gr/Ep.

Property	Material	
	Graphite/Epoxy	
	0°	90°
Tensile strength (MPa)	1378	41
Tensile modulus (GPa)	117.0–138.0	8.0–11.0
Compressive strength (MPa)	1309	N/A
Compressive modulus (GPa)	107.0–124.0	N/A
	PCD	
Knoop hardness (HK)	3400–3700	
Compressive strength (GPa)	700	
Modulus of elasticity (GPa)	827–1103	
TRS (MPa)	482–1723	
Thermal conductivity (W/m°C)	50–92	

ple 2. Edge trimming test matrix.

phite/epoxy fibre orientation		0°, 15°, 30°, 45°, 60°, 75°, 90°					
D insert	α	0°	0°	5°	5°	10°	10°
ometry	γ	7°	17°	7°	17°	7°	17°
ting speed		4, 9, and 14 m/min					
oth of cut		0.127, 0.254, and 0.381 mm					

aracteristics were recorded with a CCD Hi-8 format camera. Cutting forces were mea-
ed using a 3-dimensional circular-type strain gauge dynamometer attached to the tool
st.

Machining tests were conducted following a full factorial design of experiments, in-
porating the conditions listed in table 2. All machining was carried out under dry
ting conditions conforming to industrial specifications. Methods of analysis included
th *in-situ* and post-process measurements of machining characteristics. *In-situ* analysis
hniques consisted of optical recording of chip formation and cutting force measure-
nts along the principal and thrust directions via the 3-axis dynamometer. All aspects
chip formation from inception through chip release were recorded. Surface topography
d machined-surface characteristics were examined using profilometry and SEM. Sur-
e roughness measurements were obtained by a Federal Surfanalyzer 4000 with a 5 μm
meter diamond stylus probe, 0.8 mm cutoff length, 1.25 μm measurement interval, and
.5 mm sampling length. Microstructure of the machined surface was evaluated with
eol JSM-T330A scanning electron microscope to supplement surface roughness mea-
ements and provide a qualitative view of machining process damage. Details of this
t of study can be found in Wang *et al* (1995).

Results and discussion

this study it was found that chip formation was critically dependent on the fibre orien-
ion (Wang *et al* 1995). The change in chip formation with fibre orientation was clearly
ible by examination of the macrochip. Discontinuous chip formation was observed both
unidirectional and multidirectional laminate machining.

2a *Cutting mechanisms*: Cutting mechanisms present in machining of FRP materi-
in this study are shown schematically in figure 2. For the 0° material, chip formation
chanisms were composed of mode I loading and fracture along the fibre/matrix inter-
e, mode II loading through tool advancement, and fracture perpendicular to the fibre
ection under bending loads. In positive fibre orientations between 15° and 75°, the chip
mation mechanisms included fracture from compression-induced shear across the fibre
is combined with interfacial shearing along the fibre direction during chip advancement.
ip flow in trimming all positive orientations up to and including 90° material occurred
a plane parallel to the fibre orientation. Material removal in nearly all positive fibre
ientations, therefore, appears to be governed by the in-plane shear properties of the uni-
rectional material. Chip formation for fibre orientations greater than 75° is primarily
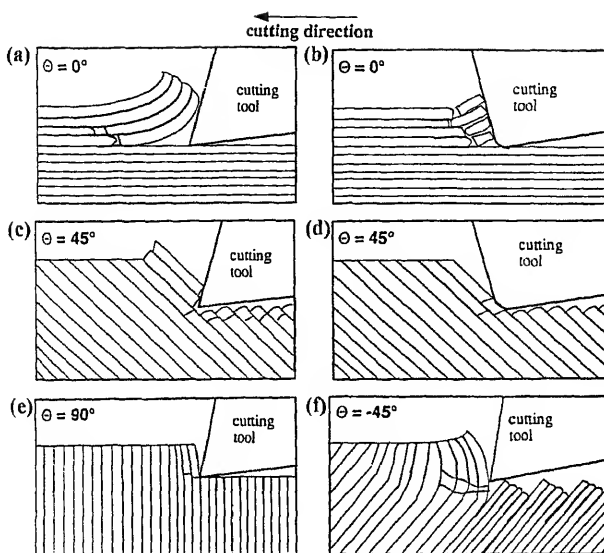
Figure 2. Cutting mechanisms in the orthogonal machining of Gr/Ep. (**a**) Delamination; (**b**) fibre buckling; (**c**) & (**d**) fibre cutting; (**e**) deformation; (**f**) shearing.

compression-induced fracture perpendicular to the fibres and interlaminar shear fracture along the fibre/matrix interface. These mechanisms are consistent with those previously documented by others (Koplev *et al* 1983; Sakuma & Seto 1983; Kaneeda 1991; Krishnamurthy *et al* 1992). Edge trimming of large fibre orientations is frequently accompanied by out-of-plane displacement and material fracture ahead of the cutting tool. Based on the results for multi-directional laminates, orthogonal cutting mechanisms for all fibre orientations were distinguished. The cutting mechanisms of 0° and 45° plies in edge trimming of multi-directional laminates were identical to those of equivalent unidirectional fibre orientations. Conventional trimming of 0° plies with the PCD tools occurred by a combination of mode I and mode II loading and subsequent fracture. Mechanisms of chip formation and material removal for 45° plies originate from shear loading. Fracture occurred perpendicular to the fibre axis within the compressive field of the contact zone, and fibre/matrix interface shear permitted chip flow. However, chip formation mechanisms in trimming 90° and −45° unidirectional material were slightly different from that for multi-directional laminates as a result of the influence of adjacent plies. When machining unidirectional composites with fibre orientations of 90° or more, extensive degrees of out-of-plane damage occurred. In contrast, support provided from adjacent plies in the multi-directional ply structure allowed for machining with relatively small degrees of structural degradation. Trimming damage associated with 90° plies of the multi-directional laminate was limited to a loss of fibre/matrix cohesion evident at the trimmed surface. Damage to the −45° plies was more extensive through shear failure along the fibre/matrix interface and fracture perpendicular to the fibre direction beneath the trimming plane. Tool geometry appeared to have no discernible effect on the degree of trimming damage. *In-situ* chip formation analysis in edge trimming the multi-directional Gr/Ep laminate revealed

similar to features of chip formation evident in trimming unidirectional material with fibre orientations greater than 90° (negative orientations). In general, chip lengths on trimming the laminate were larger than those obtained in trimming unidirectional material. Post-process characteristics of chip formation observed in edge-trimming the multi-directional laminate were also quite similar to those of unidirectional Gr/Ep, particularly in the 0° and 45° plies. Fragmented chips attributed to split failure ahead of the cutting tool were produced in the trimming of 0° plies. Very tiny chips were noted from 45° plies with chip release along the fibre direction. The fracture plane for both 0° and 45° fibre orientations was parallel to the fibre direction as in the case of 90° plies. However, contrary to trimming unidirectional Gr/Ep, tiny chips were produced in 90° laminate plies without any indication of out-of-plane displacement. Material damage was essentially eliminated due to support provided by adjacent plies in the multi-directional lay-up. Similarly, chips produced from trimming −45° plies of the laminate resulted in severe damage below the trimming plane, but the extent of damage was limited due to support provided by adjacent plies.

3.2b *Cutting forces*:   In support of chip formation dependence on fibre orientation, principal cutting and thrust forces in this study were primarily influenced by fibre orientation; operating conditions and tool geometry had very little influence. Principal cutting forces ranged from 100 N for 0° to slightly more than 800 N for 90° material. Specific cutting energy ranged from 1×10^8 to 7.8×10^8 J/m^3, significantly lower than that documented from orthogonal trimming of conventional homogeneous materials (Koplev *et al* 1983). Contrary to the mechanics of chip formation in metal cutting, the thrust force was nearly always greater than the corresponding principal force except when trimming material with fibre angles of 0°, and greater than or equal to 90°, as shown in figure 3. In general, thrust forces commonly increased with fibre orientation up to 45°, and then decreased to 90°. High thrust forces may be attributed to the elastic recovery of the fibres within the contact zone prior to fracture. The elastic energy of the deformed fibres, released when the fibres are severed, imparts a thrust force on the tool flank and is a potent source for tool wear. This analogy is supported by the fact that the thrust force in trimming 0° and 90° orientations is almost the same as they undergo the least bending-induced elastic deformation. These results are consistent with results reported by Koplev (1980), Sakuma & Seto (1983), and Inoue & Ido (1986) but not with those of Kaneeda (1991).

Features of the cutting force profiles in trimming multi-directional laminate were nearly identical to those recorded in edge trimming 90° unidirectional material. The force profiles exhibited a high frequency fluctuation in cutting force (F_c). However, tool geometry affected the magnitude of both F_c and its fluctuation in trimming of the multi-directional laminate. Compounded effects of tool rake and clearance angles on the principal and thrust cutting forces are shown in figures 4a, b; a depth of cut (t) of 0.13 mm and 9m/min cutting speed (V) were used for the simulations. Optimal rake angle (α) in terms of minimizing the principal cutting force was determined to be approximately 7° as shown in figure 4a. Thrust force, however, increased with increasing rake angle but decreased with increasing clearance angle. The optimal tool geometry for minimizing the resultant cutting force over the range of operating conditions in this study would consist of a tool with 6° ∼ 7° α and
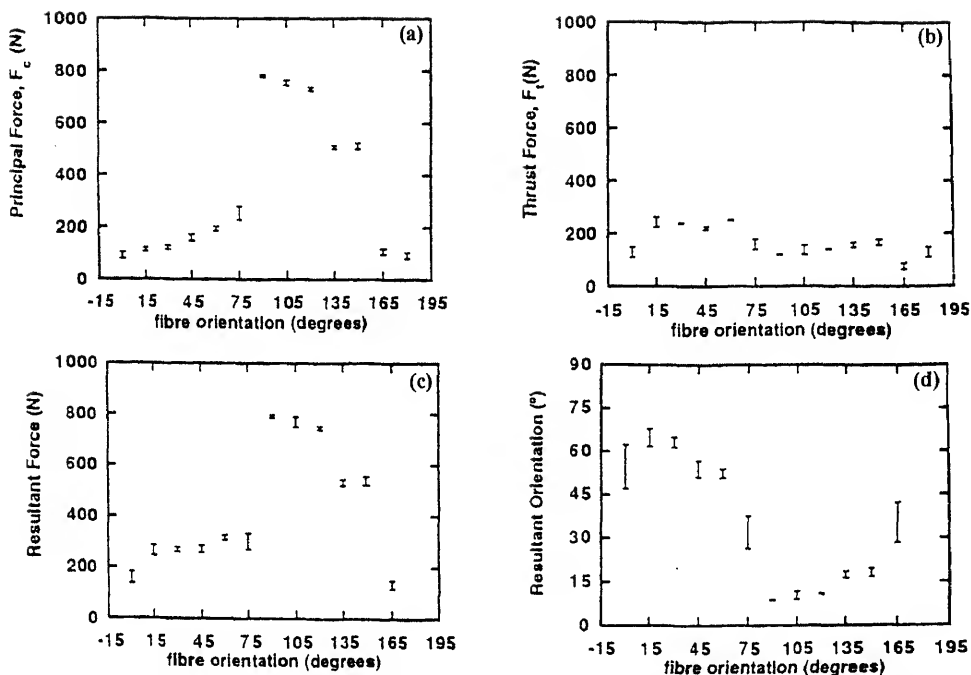
Figure 3. Cutting forces in edge-trimming with a $10°\alpha/17°\gamma$ PCD tool. Cutting conditions: 4 m/min cutting speed, 0.25 mm depth of cut. (a) Principal cutting force; (b) thrust force; (c) resultant cutting force; (d) resultant force orientation.

$17°\gamma$ angles. Increasing the clearance angle beyond $17°$ would result in lower thrust and hence resultant forces, but would ultimately result in a reduction of tool life. The optimal tool geometry to minimize cutting forces in orthogonal trimming of Gr/Ep in this study is in excellent agreement with that reported by Park (1991), which was generated from the aspect of minimizing tool wear.

Over the range of cutting speeds available from the planer/shaper machine ($V < 20$ m/min), essentially no change in cutting force was noted. However, principal cutting force increased linearly with depth of cut and in turn influenced the magnitude of the resultant cutting force and its orientation with respect to the plane of trimming. Although principal cutting force decreased with increasing rake angle, smaller resultant cutting forces were obtained with a $0°$ rake angle tool rather than positive rake angle PCD inserts due to the influence of thrust forces. In addition, smaller cutting forces were obtained for a $17°$ clearance PCD insert rather than a $7°$ clearance angle.

3.2c *Surface roughness*: Surface quality in orthogonal trimming of FRP materials is closely related to chip formation and, hence, fibre orientation. Due to the unique chip formation mechanisms of the $0°$ fibre orientation and bare fibres on the machined surface, R_a and R_y values in the longitudinal direction were usually higher than those in positive fibre orientations. The average surface roughness of fibre orientations from $15°$ to $60°$ in both longitudinal and transverse measurement directions exhibited high quality with
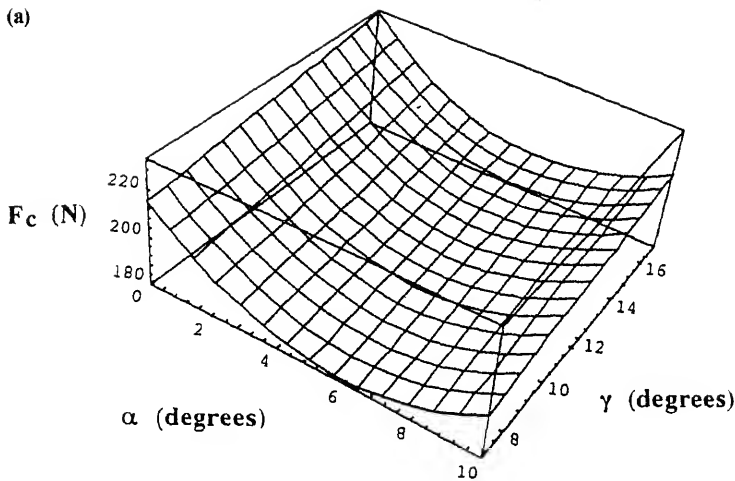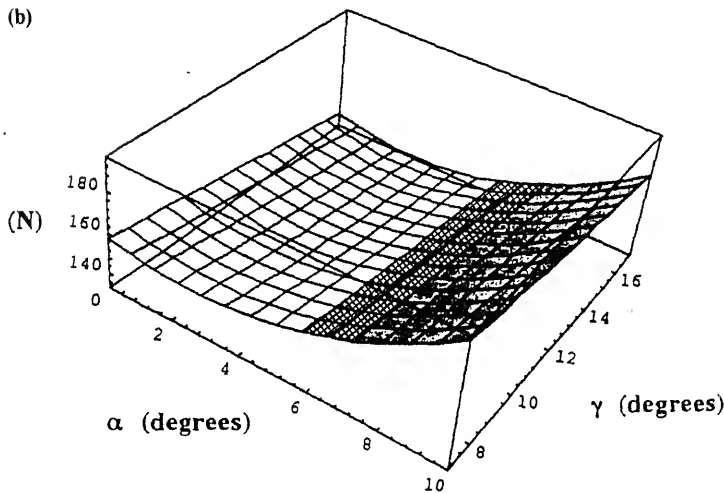
Figure 4. Effect of tool geometry on the cutting forces. Cutting conditions: 4 m/min cutting speed, 0.13 mm depth of cut. (**a**) Principal cutting force; (**b**) thrust force.

ghness between 1 and 1.5 μm. Very few changes were noted in roughness measurements ween fibre orientations of 15° and 60°. Surface texture measurements were not obtained positive orientations greater than 60° or negative orientations due to severe cutting nage as previously discussed. Peak-to-valley height (R_y) values in figure 5 exhibited ds similar to longitudinal and traverse direction R_a measurements though with a more sitive response to fibre orientation.

urface roughness measurements for laminate composites were significantly influenced the variation in ply angles and measurement direction. Matrix smearing was dominant 45° and 90° fibre plies, and bare fibres were usually observed on the 0° ply. The domi-t feature of −45° plies was the "in-depth" damage due to fibre pullout and intralaminar cking (delamination) which induced deep valleys in profile measurements taken trans-se to the trimming direction. Longitudinal measurements taken parallel to the cutting
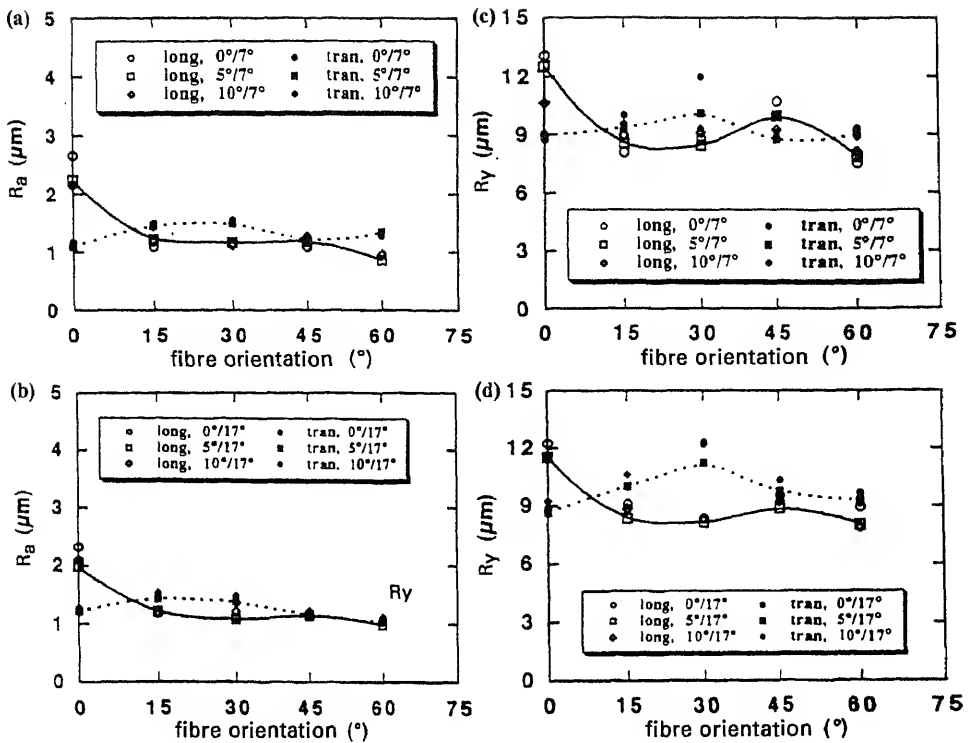
Figure 5. Tool-geometry influence on surface texture. Cutting conditions: 9 m/min cutting speed, 0.25 mm depth of cut. Average roughness with 7°γ (**a**); and 17°γ (**b**) average peak-to-valley height with 7°γ (**c**) and 17°γ (**d**).

direction did not indicate the extent of trimming damage, even when acquired on the −45° plies as shown in figure 6. Therefore, average surface roughness (R_a) and maximum height roughness (R_y) alone do not appear capable of indicating the "true" surface quality of a trimmed FRP without additional methods of observation. A combination of standard roughness parameters and a statistical representation of profiles taken perpendicular to the feed direction as shown in figure 6 may be more appropriate for a qualitative inspection.

3.3   *Conclusions*

Orthogonal cutting mechanisms in the edge-trimming of graphite/epoxy laminates with polycrystalline diamond tools were studied. Chip formation, cutting force, and surface morphology were evaluated with respect to tool geometry, process conditions, and ply distribution in the laminate. As with the trimming of unidirectional panels, discontinuous chips were observed with primary dependence on fibre orientation. Cutting mechanisms for 0° and 45° plies were identical to those in the trimming of unidirectional material. However, chip-formation mechanisms in trimming 90° and −45° plies of the multi-directional
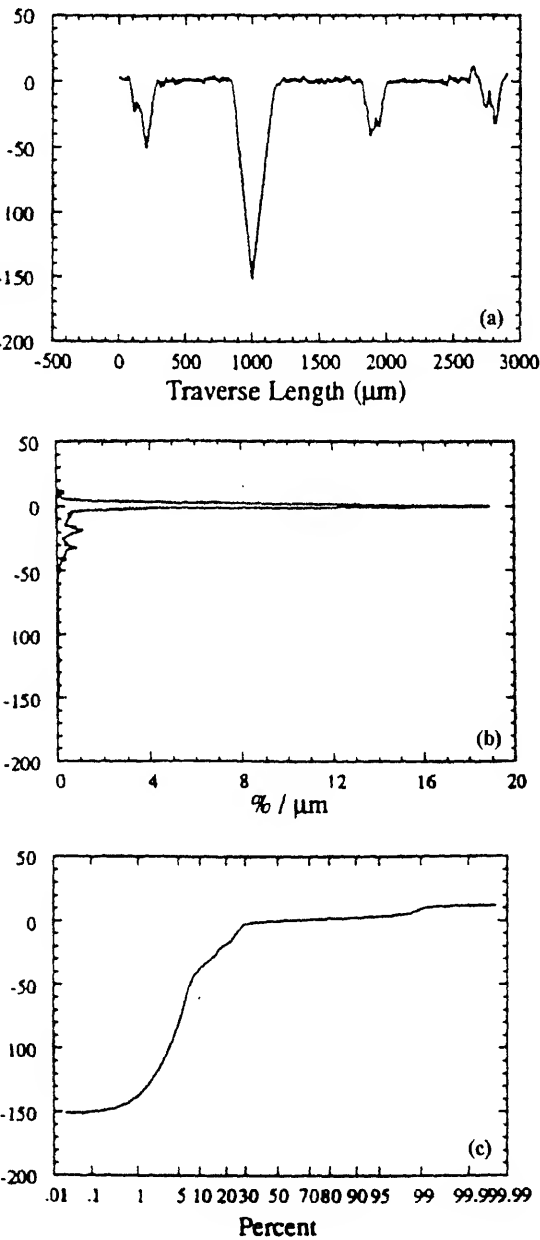
Figure 6. Transverse profile of the Gr/Ep laminate. Cutting conditions: 4 m/min cutting speed, 0.25 mm depth of cut, 0°α/17°γ tool. (a) Profile height; (b) probability density of profile height; (c) cumulative height distribution.

4.1 *Materials and procedures*

Graphite/epoxy laminate material with 3501-6 resin and IM-6 fibres was used for all experiments in this preliminary investigation. Compression specimens were composed of plies of unidirectional tape (0° plies), plain-weave fabric (0°/90°), and bias-weave fabric (45°/−45°). Flexure specimens consisted of unidirectional tape laminated with combinations of 0°, 90°, 45°, and −45° plies. Panels used for both phases of testing were hand laminated with lay-up indicative of that used in aerospace structures.

Compression specimens were acquired from a bulk laminate panel using either diamond abrasive cutters or AWJ machining according to a standard aerospace coupon design (Colligan 1993). All specimens were first rough-machined with a #30 grit diamond abrasive cutter. A test section of the dog-bone specimen was obtained using either diamond abrasive cutters or the abrasive waterjet incorporating various parametric conditions which provided a range in edge quality received from each machining technique. Flexure and impact specimens were obtained using three manufacturing techniques, namely traditional trimming with polycrystalline diamond (PCD) cutting tools, AWJ machining, and diamond-saw machining with a #220 garnet diamond-impregnated slot saw. Cutting conditions used in obtaining the compression and flexure specimens are available in Colligan (1993) and Arola & Ramulu (1994) respectively. Surface profiles of the machined specimens were obtained using a stylus profilometer according to ANSI B46.1-1985. Surface topography of the machined specimens was analysed from profile data using standard roughness parameters, statistical methods, and random process methods in addition to visual techniques. Details concerning analysis procedures are provided in the aforementioned references.

An MTS Universal Test Machine was used for compression tests with a stroke rate of 0.8 mm/min, hydraulic grips, and buckling constraint. An extensometer was used to record strain data. Flexure and impact specimens were loaded to failure according to ASTM standard D790M (ASTM 1986). Four-point and three-point flexure configurations were used with 16-to-1 span-to-depth ratio. Machined edges of the flexure and impact specimens were parallel to the plane of applied load. Testing performed in both modes of loading was performed at ambient temperature and humidity. Ultimate load and associated failure phenomena were correlated with qualitative data of the machined specimens to study the influence of manufacturing and surface quality on mechanical performance under compression and bending loads.

4.2 *Results and discussion*

Study of machining influence on the performance of Gr/Ep subjected to compression and pure bending under four-point loading was conducted (Colligan 1993; Arola & Ramulu 1994). Prior to conducting strength analysis, topographies of the machined surfaces obtained with each method of machining were analysed using profilometry. Typical profiles obtained from the machined surface of diamond-saw, PCD-trimmed, and AWJ-machined specimens as well as the profile probability density, and cumulative height distribution are shown in figures 7a–c respectively. Profiles in figure 7a were obtained perpendicular to
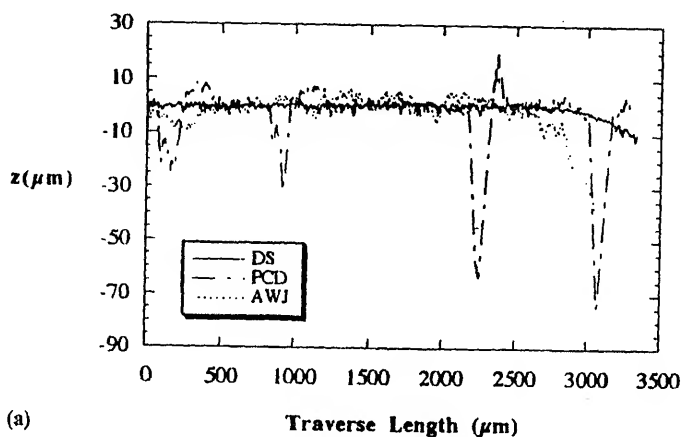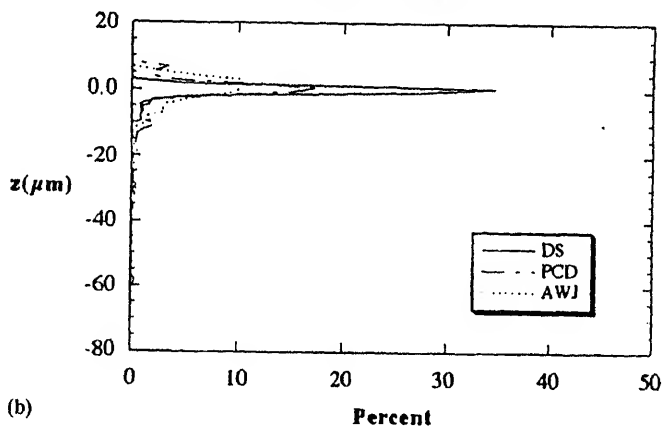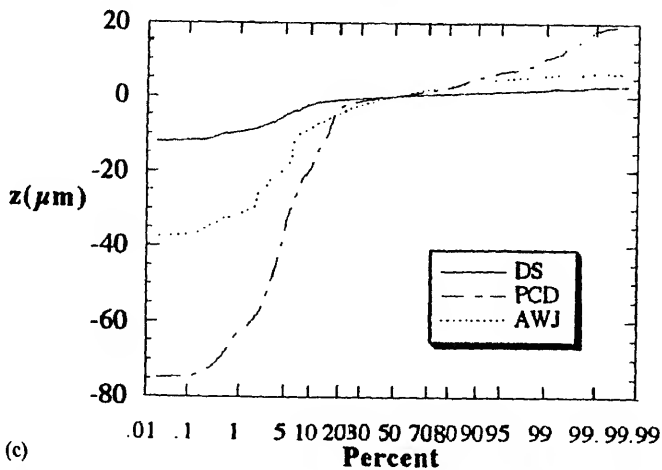
Figure 7. Flexure specimen transverse profilometry. (a) Profile height; (b) profile height
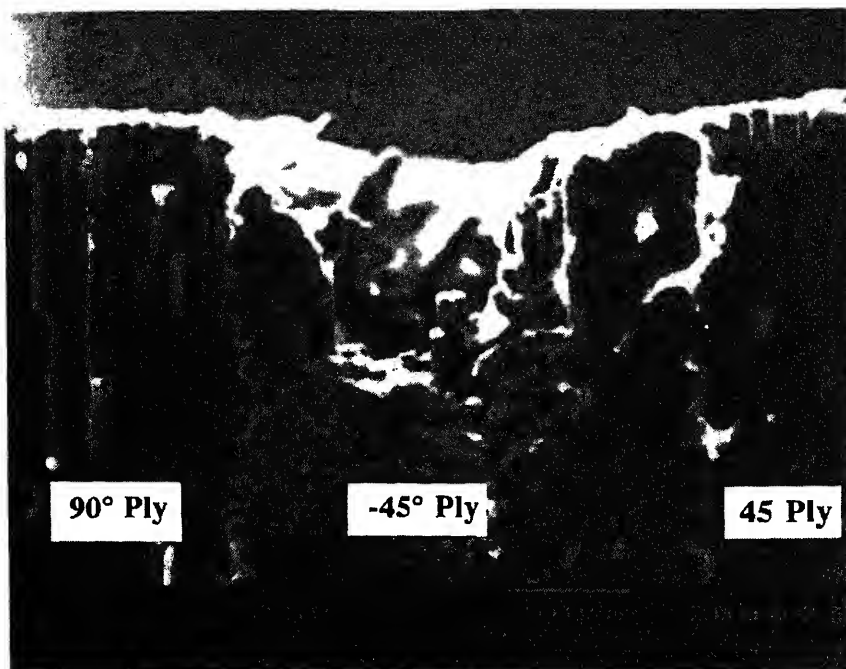
Figure 8. Damage to 45° ply of PCD edge-trimmed Gr/Ep laminate.

the major axis (feed direction). In general, all three techniques produced relatively high quality machined surfaces. However, although the average surface roughness (R_a) of the PCD specimens ranged between 1 and 2 μm, extensive fibre pullout occurred with the −45° plies as evident in figure 8. Plies in this orientation exhibit fibre direction with slope opposite to the cutting direction and commonly fail along the fibre/matrix interface during traditional edge-trimming. Compression specimens obtained with the AWJ and diamond abrasive cutters were machined at process conditions that provided a wide range in surface quality. Average surface roughness of specimens fabricated with each technique ranged from 5 to 25 μm R_a.

Following the qualitative inspection of the machined surfaces, compression, flexural and impact testing was conducted. While recording the stress strain response of compression specimens, it was noted that the AWJ-machined specimen loading response appeared to be a function of surface quality. Typical stress strain curves for three AWJ machined specimens of differing quality are shown in figure 9. Initiation and propagation of failure occurred first in the specimens with high surface roughness and exit ply delamination. Typical load displacement histories for PCD-trimmed, diamond-saw and AWJ-machined flexure specimens are shown in figure 10. Load deflection curves obtained during application of the bending load consistently suggested that permanent damage, i.e. failure initiation and propagation, occurred first in PCD-trimmed specimens.
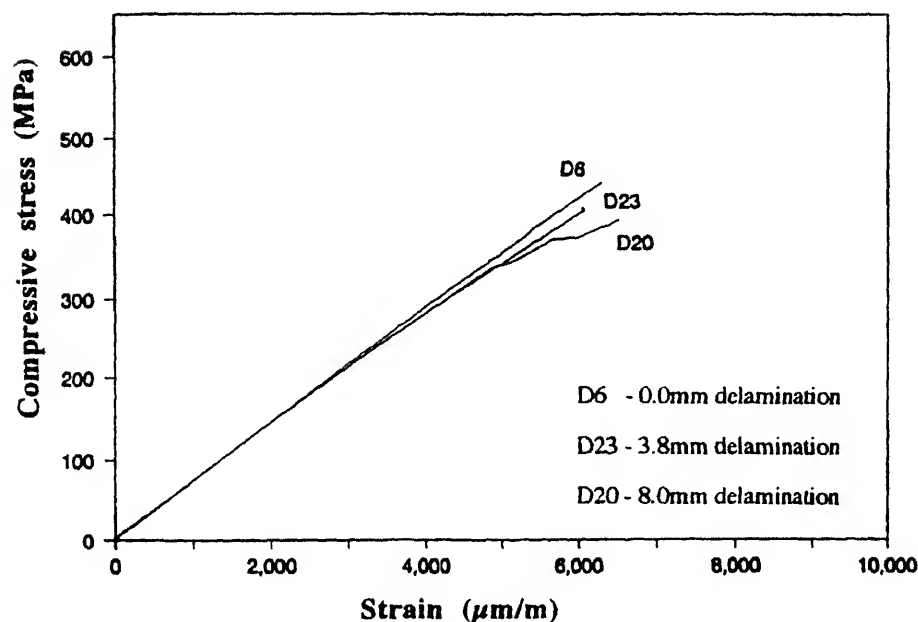
Figure 9. Stress/strain history of AWJ-machined Gr/Ep compression specimens.

Subsequent analysis showed that surface roughness resulting from trimming with dia-
nd abrasive cutters had essentially no influence on the ultimate compression strength.
wever, although surface roughness of AWJ specimens had notable but limited influence
the ultimate compression strength within comparable roughness of diamond abrasive
cimens, exit ply delamination had significant influence. The relationship between sur-
e roughness and ultimate compression strength from both processing techniques are
wn in figures 11a–b. The correlation between exit ply delamination resulting from
/J trimming and ultimate compression strength is also shown in figure 11c.

Similar to compression specimen analysis, characteristics of the flexure specimen ma-
ned surface indicative of each fabrication method were quite different, but no distinction
ltimate strength could be made from average surface roughness. Mean and characteris-
strengths calculated for the three specimen batches from Weibull statistics were nearly
uivalent.

Typical preliminary impact test results (Arola & Ramulu 1995) are shown in figure 12. A
ical load and the load-line displacement profile for a diamond-saw specimen is shown
figure 12a. The features of the impact event of interest include the peak load, peak
cture load, and the total energy to failure. Representative load and load-line displace-
nt profiles for PCD and AWJ specimens are shown in figure 12b–d. Unfortunately, the
namic sensors available to record specimen displacement in this study had a travel limit
5 mm. Therefore, the energy to complete failure, represented by the integral of the total
remental load displacement curve, was not available. Hence, the energy calculated to
limit of the displacement sensors is essentially the total energy to fracture. Averaged
ues for the impact properties of 100 specimens are listed in table 3. Only minimal dif-
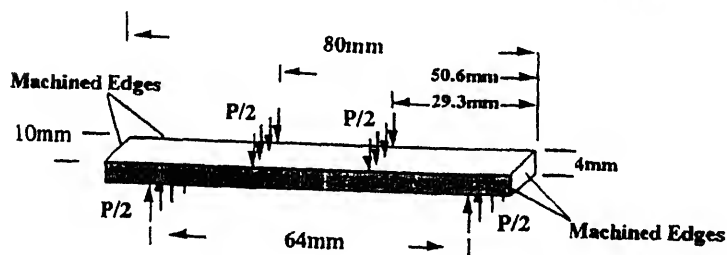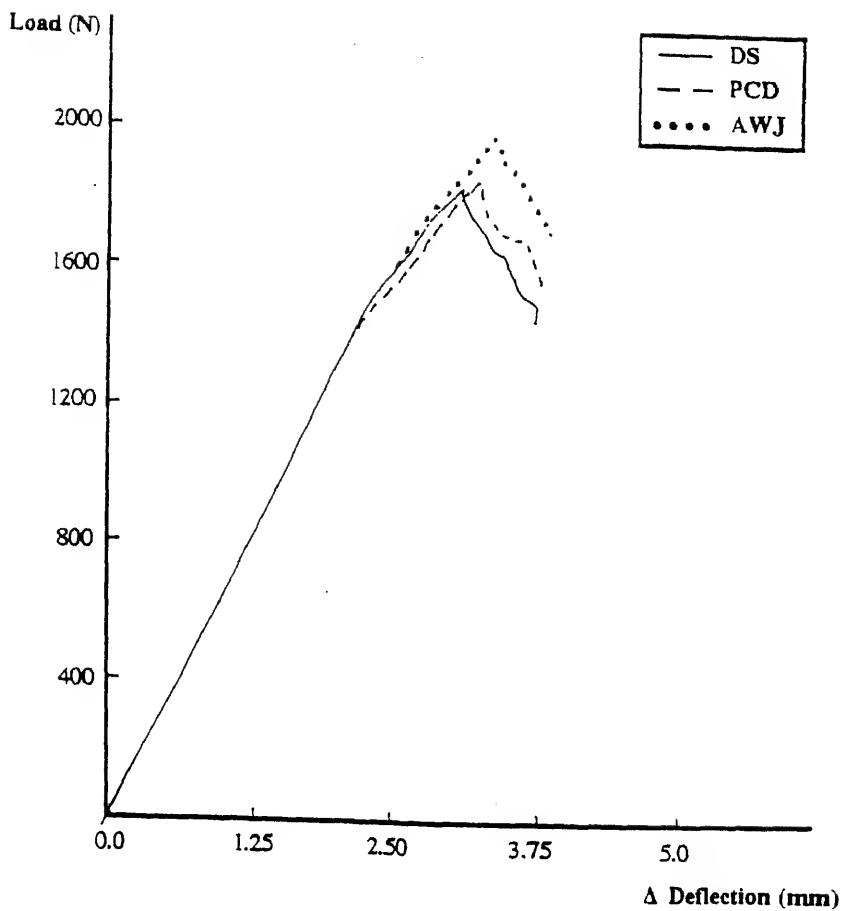
Figure 10. Flexure specimen geometry and loading configuration.
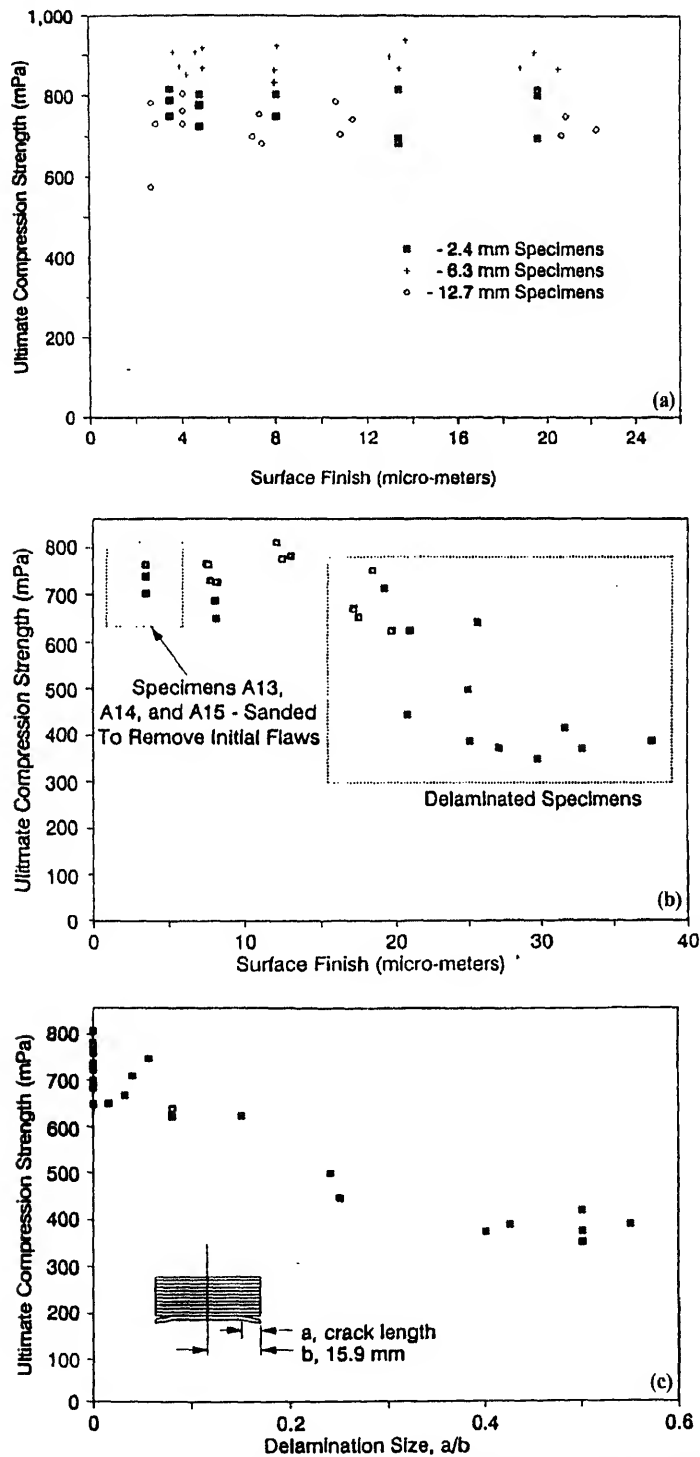
Figure 11. Influence of processing on the ultimate compression strength. (a) Machined with diamond abrasive cutters; (b) AWJ-machined; (c) AWJ compression strength vs. exit
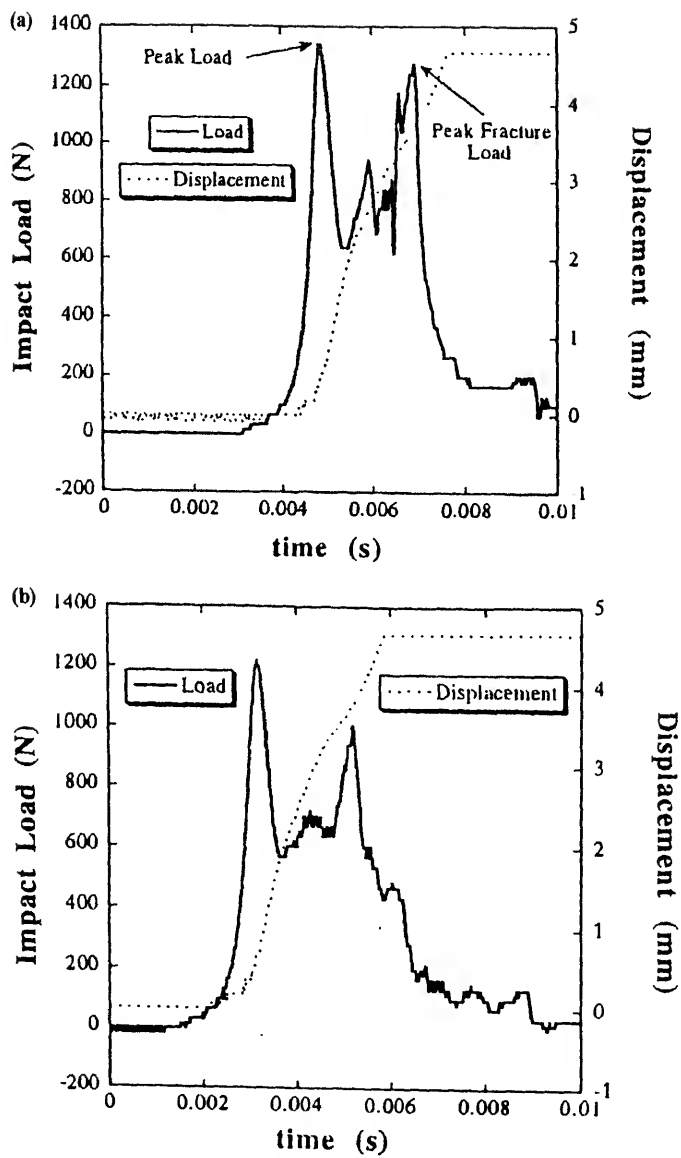
Figure 12.   (a and b. Caption on facing page.)
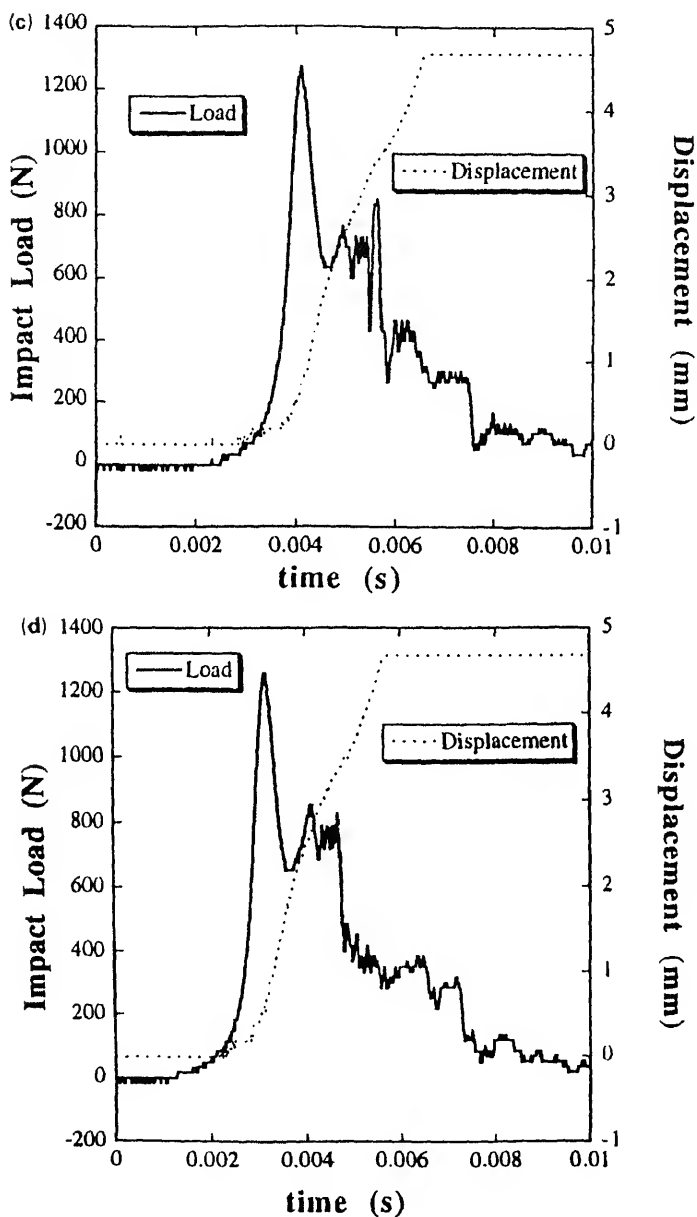
Figure 12. Typical recorded load, load-line displacement, (a) diamond-saw, (b) PCD, (c) AWJ (a) (no. 50 garnet) . (d) AWJ (c) (no. 150 garnet).

Table 3.   Impact loading failure parameters.

Machining type	Peak load (N)	Peak fracture load (N)	Energy to fracture (N·m)
Diamond-saw	1285.8 ± 60.8	1057.9 ± 207.9	3.43 ± 0.42
Planer (PCD)	1235.1 ± 29.8	1028.0 ± 155.1	3.51 ± 0.15
AWJ (a)	1215.0 ± 67.6	817.7 ± 72.0	3.04 ± 0.15
AWJ (b)	1258.1 ± 44.6	904.5 ± 107.5	3.24 ± 0.14
AWJ (c)	1262.6 ± 48.1	958.4 ± 206.9	3.29 ± 0.26

in peak fracture load and the energy dissipated varied with the method of machining. As expected, the diamond-saw had the highest peak load to failure and energy to fracture.

Scanning electron microscope analysis of failed specimens suggested that features resulting from traditional trimming with PCD tools may have given rise to unique failure propagation. Extensive macro-cracking along the $-45°$ plies occurred in PCD-trimmed specimens extending through the entire specimen length. Conversely, diamond-saw and AWJ-machined specimens exhibited much higher post-failure integrity with far fewer macro-cracks. Results from compression, flexure and impact tests suggest that surface finish alone does not degrade a laminate. Delaminations incurred during trimming control residual strength.

4.3   *Conclusion*

Machining-induced surface features discussed in this study had limited influence on bulk strength of the Gr/Ep material but may affect the impact or fatigue strength where flaws in material structure become more important. Preliminary results of impact tests indeed revealed the effect of edge-finishing on the peakloads to failure. In these loading configurations, fibre pullout and fibre matrix debonding, as well as the high values of surface roughness, cause flaw propagation and accelerated failure. Further study of the effects of machining-induced damage on composite material performance will address this issue. Studies by Howarth & Strong (1990) combined with results from this preliminary investigation confirm that aspects of net-shape manufacturing, including the particular technique used and the resulting cut characteristics, may influence the structural integrity of FRP's. However, the mechanisms responsible for reduction in mechanical performance are not clearly understood.

5.   **Summary**

Investigation of machining mechanisms and machining induced edge effects on the structural integrity of FRP materials was conducted. Consistent with previous reports, results from compression, flexural and impact testing to failure suggest that particular features of the machined surface may not only affect the bulk strength, but also the initiation and propagation of failure. Therefore, for predicting the future applications of FRP materials, based on the current state of applicable net-shape manufacturing technology, additional investigations are necessary to thoroughly understand the effects of the methods of manufacturing on the structural integrity of FRP materials.

References

Abrate S, Walton D A 1992a Machining of composite materials. Part I: Traditional methods. *Compos. Manuf.* 3: 75–83

Abrate S, Walton D 1992b Machining of composite materials. Part II: Non-traditional methods. *Compos. Manuf.* 3: 85–94

Arola D, Ramulu M 1994 Machining induced surface texture effects on the flexural properties of a graphite/epoxy laminate. *Composites* 25: 822–833

Arola D, Ramulu M 1995 Manufacturing effects on the impact properties of graphite/epoxy composite. *Proceedings of the 10th Annual Technology Conference of American Society of Composites* (in press)

ASTM 1986 *Standard test methods for the flexural properties of unreinforced and reinforced plastics and electrical insulating materials.* ASTM D790M

Colligan K 1993 *Machined edge effects on the compression strength of graphite/epoxy.* MS thesis, University of Washington, Seattle

Colligan K, Ramulu M 1992 The effect of edge trimming on composite surface plies. *Manuf. Rev.* 5: 274–283

Ghasemi Nejjad M N, Chou T-W 1990 Compression behaviour of woven carbon fibre-reinforced epoxy composite with molded in and drilled holes. *Composites* 21: 33–40

Griswold N C, Massarweh W A, Hough C L 1989 Morphological structures in analysis and inspection of hole quality in fiber composites. *J. Testing Eval.* 17: 281–286

Ho-Cheng H, Dharan C K H 1990 Delamination during drilling in composite laminates. *J. Eng. Ind.* 112: 39–47

Howarth S G, Strong A B 1990 Edge effects with waterjet and laser beam cutting of advanced composite materials. In *35th International SAMPE Symposium*, pp 1685–1697

Inoue H, Ido M 1986 Study on the cutting mechanism of GFRP. In *Proceedings of Int. Symp. on Composite Materials and Structures*, Beijing, pp 1110–1115

Jamil S J, Chambers A R 1991 Evaluation of surface quality of drilled holes in composite materials after high speed drilling. In *Advance machining for quality and productivity* (incorporating the 2nd International Conference on the Behavior of Materials in Machining, 1991), pp 70–76

Kaneeda T 1991 CFRP Cutting mechanism. *Trans. North Am. Manuf. Res. Inst. SME* 19: 216–221

Kinkaid R 1988 Quality holes in composites with PCD cutting tools. *Cutting Tool Eng.* 40: 50–52

König W, Wulf C, Graf P, Willersceid H 1985 Machining of fiber reinforced plastics. *Ann. CIRP* 34: 537–547

Koplev A 1980 Cutting of CFRP with single edge tools. In *Proc. of 3rd Int. Conf. on Composite Materials*, Paris, pp 1597–1605

Lin H J, Lee Y J 1992 Strength of composite laminates with continuous fiber around a circular hole. *Compos. Struct.* 21: 155–162

McCarty J E 1993 Design and cost viability of composites in commercial aircraft. *Composites* 24: 361–365

Mehta M, Reinhart T J, Soni A H 1992 Effect of fastener hole drilling anomalies on structural integrity of PMR-15/Gr composite laminates. In *Machining of composites* (Columbus, OH: ASM) pp 113–126

Miller R K 1991 Waterjet cutting: Technology and industrial applications (Liburn, GA: Fairmont)

Mortimer J 1987 New technology brings quality to manufacture. *Ind. Robot* 14: 103–104

Park J 1991 *Study on the effect of PCD tool geometry on the machining of graphite/epoxy composite materials.* MS thesis, University of Washington, Seattle

Pengra J J, Wood R E 1980 The influence of hole quality on graphite/epoxy composite laminates. In *Proceedings of the AIAA/ASME/ASCE/AHS 21st Structures, Structural Dynamics and Materials Conference* AIAA Paper 80–0777, pp 687–693

Philips J L, Parker R T 1987 Fastener hole considerations. In *Engineered materials handbook, Composites* (Columbus, OH: ASM) pp 712–715

Pipes R B, Pagano N J 1970 Interlaminar stresses in composite laminates under uniform axial extension. *J. Compos. Mater.* October: 538–548

Ramulu M, Arola D 1993 Water jet and abrasive water jet cutting of unidirectional graphite/epoxy composite. *Composites* 24: 299–308

Ramulu M, Arola D 1994 The influence of abrasive waterjet cutting conditions on the surface quality of graphite/epoxy laminates. *Int. J. Mach. Tools Manuf.* 34: 295–314

Ramulu M, Wern C W, Garbini J L 1993 Effect of fiber direction on surface roughness measurements of machined graphite/epoxy composite. *Compos. Manuf.* 4: 39–51

Riggs J P 1984 Emerging non-metallic structural materials used for airframes and other demanding applications. *Mater. Soc.* 8: 351–376

Sadat A B 1988 Machining of graphite/epoxy composite material. *SAMPE Q.* 19: 1–4

Sakuma K, Seto M 1983 Tool wear in cutting glass-fiber-reinforced plastics. *Bull. JSME* 26: 1420–1427

Tagliaferri V, Di Ilio A, Crivelli Visconti I 1985 Laser cutting of fibre reinforced polyesters. *Composites* 16: 317–325

Tagliaferri V, Caprino G, Diterlizzi A 1990 Effect of drilling parameters on the finish and mechanical properties of GFRP composites. *Int. J. Mach. Tools Manuf.* 30: 77–84

Wang D H 1993 *Machining characteristics of graphite/epoxy composites.* Ph D thesis, University of Washington, Seattle

Wang D H, Ramulu M, Wern C W 1992 Orthogonal cutting characteristics of graphite/epoxy composite material. *Trans. NAMRI/SME* 20: 159–165

Wang D H, Ramulu M, Arola D 1995a Orthogonal cutting mechanisms of graphite/epoxy composite, Part I: Unidirectional laminate. *Int. J. Mach. Tools Manuf.* 35: 1623–1638

Wang D H, Ramulu M, Arola D 1995b Orthogonal cutting mechanisms of graphite/epoxy composite, Part II: Multi-directional laminate. *Int. J. Mach. Tools Manuf.* 35: 1639–1648

Wern C W 1991 *Surface characteristics of machined graphite/epoxy composites.* M S thesis, University of Washington, Seattle

Wood R E 1978 Graphite/epoxy composite hole quality investigation. In *10th National SAMPE Technical Conference, Materials Synergisms* 10: 636–650

You S-S, Chou T-W 1988 Strength of woven-fabric composites with drilled and molded holes. In *8th Conference on Composite Materials: Testing and Design*, ASTM STP 972, pp 423–437

Machining and surface finishing of brittle solids

S CHANDRASEKAR and T N FARRIS

School of Engineering, Purdue University, West Lafayette, IN 47907-1282, USA
e-mail: chandy@ecn.purdue.edu

Abstract. Ceramic materials are finished primarily by abrasive machining processes such as grinding, lapping, and polishing. In grinding, the abrasives typically are bonded in a grinding wheel and brought into contact with the ceramic surface at relatively high sliding speeds. In lapping and polishing, the ceramic is pressed against a polishing block with the abrasives suspended in between them in the form of a slurry. The material removal process here resembles three-body wear. In all these processes, the mechanical action of the abrasive can be thought of as the repeated application of relatively sharp sliding indenters to the ceramic surface. Under these conditions, a small number of mechanisms dominate the material removal process. These are brittle fracture due to crack systems oriented both parallel (lateral) and perpendicular (radial/median) to the free surface, ductile cutting with the formation of thin ribbon-like chips, and chemically assisted wear in the presence of a reactant that is enhanced by the mechanical action (tribochemical reaction). The relative role of each of these mechanisms in a particular finishing process can be related to the load applied to an abrasive particle, the sliding speed of the particle, and the presence of a chemical reactant. These wear mechanisms also cause damage to the near ceramic surface in the form of microcracking, residual stress, plastic deformation, and surface roughness which together determine the strength and performance of the finished component. A complete understanding of the wear mechanisms leading to material removal would allow for the design of efficient machining processes for producing ceramic surfaces of high quality.

Keywords. Ceramic surfaces; abrasive machining processes; surface finishing; wear mechanisms.

1. Introduction

The thrust towards improved efficiency in gas turbines and internal combustion engines, the improved performance of wear resistant components, and the unique electrical and magnetic property requirements in electronic devices and sensors have focused attention

ceramics over other materials include high hardness and strength at elevated temperatures, chemical stability, low friction, and high wear resistance. However, those properties that give ceramics superior wear resistance also make them difficult to machine. In addition, the limited ductility of these ceramics makes forming methods that rely on extensive plastic deformation useful only for ceramics in the green state. Thus, extensive machining is required for manufacture of complex shapes with high quality surfaces. Machining costs can constitute up to 80% of the cost of the manufactured component. A detailed understanding of the wear mechanisms underlying the machining of ceramics and the damage that they leave behind should allow for more economical manufacture of reliable advanced ceramic components.

Recent authors have reviewed several aspects of wear of ceramics including Braza *et al* (1989) who overviewed its relationship to contact fatigue, Larsen-Basse (1994) who compared and contrasted wear of ceramics with much that is already known about cemented carbides and cermets, and Jahanmir & Dong (1994) who give examples of maps of wear regimes for relevant contact pressure and temperature. The following reviews mechanical aspects of wear associated with machining.

2. Material removal processes

While turning and milling are used extensively in machining of metals, they are not efficient for fully densified ceramics due to rapid tool wear and large amounts of surface damage. Diamond turning can be used for machining ceramics in the green state but finish machining is still required on the densified ceramic. Thus, surface finishing of ceramics is primarily accomplished by abrasive finishing processes such as diamond grinding, lapping, and polishing. These processes are required to meet the stringent tolerance and surface finish requirements in structural and electronic ceramics (e.g. < 0.05 μm rms in magnetic recording heads, silicon wafers, face seals, and bearings). The abrasive processes remove material mechanically and introduce damage on the surface of ceramics (Marshall *et al* 1983). This damage is usually in the form of residual stresses and cracks which have a major influence on the mechanical properties and integrity of the machined ceramic surface. Moreover, plastic deformation and subsequent residual stresses induced by surface finishing alter electro-magnetic properties such as permeability, resistivity, and refractive index of the surface in electronic ceramics (Stokes 1972) causing deterioration of their performance in electronic devices.

In general, the material removal rates (MRR) in grinding are higher than the MRR in lapping and polishing. The higher MRR in the initial rough grinding operation leads to surface damage in the form of microcracks which may extend as deep as 100 μm into the surface. Rough grinding is followed in turn by fine grinding, lapping and polishing in which this damage is removed to varying degrees. Furthermore, the lapping and polishing operations may also leave damage behind on the surface in the form of residual stresses and severely plastically deformed layers.

Grinding of ceramics is accomplished primarily with grinding wheels containing diamond abrasive grits (Subramanian 1988). The diamonds are fixed to the wheel through relatively compliant resin bonds or stiff vitrified bonds. The grits are statistically distributed over a range of sizes with an average size of $\sim 100\,\mu m$ in wheels designated as coarse and $\sim 5\,\mu m$ in wheels designated as fine. Typical process parameters used in grinding include wheel surface speeds of 25 to 50 m/s, depths of cut of 0.5 to 30 μm, and table traversal speeds of ~ 20 mm/s. These parameters lead to an MRR per unit wheel width of the order of 0.1 to 1 mm^3/mm/s and normal grinding forces per unit wheel width of 5 to 100 N/mm. For a given depth of cut, the grinding force typically increases as the abrasive particle size in the wheel decreases. Here the depth of cut refers to the depth of material removed in a single traversal of the wheel across the ceramic surface. However, the depth removed by a single particle is much smaller than this and varies over the length of contact between wheel and workpiece. The length of contact is $l = \sqrt{2Rd}$ (neglecting wheel deflection) where R is the radius of the wheel and d is the depth of cut; typical contact lengths are in the range of 1 to 2 mm.

The force applied to the grinding wheel produces wheel and machine deflections so that the actual length of contact varies from that calculated using geometry (Hucker *et al* 1994). The wheel deflection is due to the localized contacts between the stiff diamond abrasive particles (embedded in a compliant bond) and the workpiece as well as a global deflection due to the stress distribution associated with transmission of the total contact force through the wheel. A definitive calculation of the number of active abrasive particles, i.e. the number of particles actually engaged in the cutting action, and the distribution of forces on these particles is as yet unavailable. However, the statistical distribution of particles on the wheel surface suggests that many of the particles have depths of cuts much less than but of the order of the grinding depth of cut. Furthermore, the contact pressure between a single particle and the ceramic surface is very high and approximately equal to the ceramic hardness. This is to be contrasted with the average pressure across the wheel-ceramic interface estimated from force measurements, which is well within the elastic range.

The large relative sliding velocity produces high temperatures at the abrasive-workpiece interface. For typical grinding conditions, infrared radiometric measurements of the peak temperatures have shown them to be as high as 1300°C (Hebbar *et al* 1992). These values are consistent with measurement and analysis of temperatures in single point grinding under similar grinding conditions. Further substantiation of high grinding temperatures is provided by the presence of spherical particles in grinding swarf while grinding hardened steels (Lu *et al* 1992). The grinding temperatures are localized near the surface and the resulting thermal gradients generate thermal stresses which are also important in understanding the grinding process.

The localized nature (i.e. local to the single abrasive particle-ceramic contact) of the contact stresses and temperatures during grinding, and a variety of observations pertaining to deformation and stresses on machined surfaces, strongly suggest that to understand the material removal mechanism during ceramic grinding, it would be useful to analyse the sliding indentation of a ceramic surface by a hard particle under depths of cut and sliding velocities occurring in grinding. Indeed, this view underlies much of the following discussion.

Lapping and polishing of ceramics are carried out by placing a slurry of abrasive particles in a liquid vehicle between the specimen and a hard block (lapping) or a soft pad (polishing).

The block is loaded against the workpiece by hydraulic or mechanically applied pressure and rotated at slow speed. The particles roll and slide across the ceramic surface so that the wear resembles three-body wear as opposed to the two-body wear in grinding. Polishing is usually performed after initial grinding has been used to generate the workpiece shape; its primary purpose being the generation of smooth surfaces. A hard lapping block is used when stringent tolerances on the flatness of the workpiece are required, while polishing generates a much smoother surface since many of the abrasive particles are embedded in the soft lapping block. Diamond particles are used extensively but softer abrasives such as Al_2O_3, SiC, and cerium oxide are also widely used. As in grinding, the particles are statistically distributed over a range of sizes with the average particle size ranging from 0.05–70 μm.

The mean sliding velocity between the abrasive and the ceramic in lapping and polishing is of the order of 0.5 m/s or less which is two orders of magnitude less than that in grinding. In contrast to grinding, the sliding induced temperatures are thought to be insignificant in lapping and polishing. The forces on individual abrasive particles vary, with the total force applied to the block or pad being a process variable. In general, the surface roughness (R_a) of the finished surface increases with increasing lapping pressure. Typical removal rates in lapping and polishing range from $0.001 - 1$ mm^3/s which is less than that observed in grinding for a typical grinding wheel of 5 mm in width. The smaller MRR suggests that less damage is left behind on the surface from lapping and polishing than from grinding. As in grinding, the material removal can be viewed as due to sliding indentation for the particles that slide, and due to quasi-static indentation for particles that roll. This idealization is consistent with microscopic observations of lapped and polished surfaces, and wear particles formed by these processes (Chauhan et al 1993).

Recently, a model for the distribution of abrasive particle forces during lapping based on statistical size distribution of the particles has been developed (Chauhan et al 1993). In that paper, the required compliance of the particle workpiece contact is calculated by assuming that it is the same as that for an indentation with a conical indenter. An interesting result from the calculation is that approximately only 1 out of 10^5 particles is actively engaged in material removal at a given time. Further, by assuming that the surface roughness of the finished surface is related to the depth of the plastic zone produced by a particle, the surface roughness could be predicted from the properties of the abrasive particles and the workpiece, and the lapping pressure. The mean surface roughness, R_a, is found to be related to the average particle force while the peak-to-valley surface roughness, R_t, is related to the maximum force applied by a particle. The predictions of this model are, for most cases, in excellent agreement with experimental observations from lapping and polishing of Al_2O_3, soda-lime glass, and Ni-Zn ferrite using SiC abrasive slurries of different particle sizes. A similar calculation has not been carried out for grinding in part because of difficulties associated with measurement of the surface profile of the grinding wheel and lack of characterization of the statistical distribution of particle sizes on the wheel surface.

3. Material removal mechanisms

A glimpse into the wear processes that cause material removal during the abrasive machining of ceramics is provided by optical and electron microscopy of machined ceramic
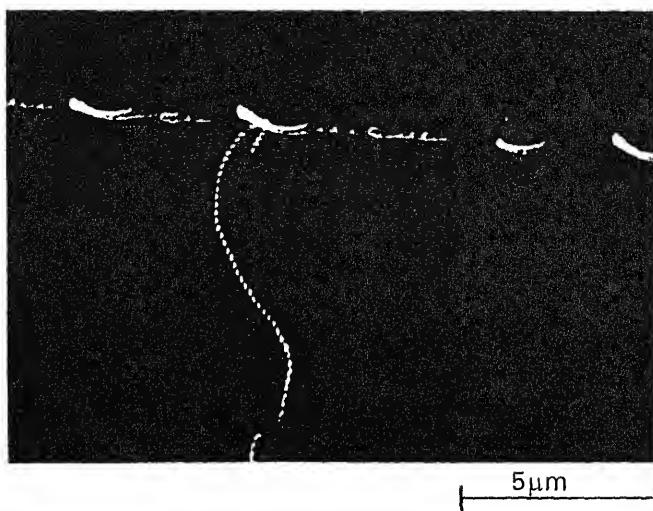
Figure 1. SEM micrograph of ribbon-shaped polishing chips from soda-lime glass.

surfaces and machining chips. Such microscopic observations have shown that material removal occurs as a consequence of one or more of the following: lateral cracks breaking open onto the surface; gross fracture due to grain pull-out; median/radial cracks intersecting with each other; and plastic micro-cutting through the formation of chips as in single-point turning of metals. These observations have also been confirmed through electron microscopic studies of chips and wear particles formed by machining. The dominance of a given mechanism is closely related to the loads applied by individual abrasives on the ceramic surface during machining. Typically, when the externally applied loads transmitted by the abrasives are small, plastic micro-cutting or indentation (with upward displacement of material around the indent) mechanisms are found to dominate. This is particularly so when polishing at low loads and/or with a flexible, soft polishing cloth (polishing block) or in the so-called "ductile regime" mode of grinding. The surfaces produced by these modes of material removal/displacement are characterized by their extreme degree of smoothness. Also, the plastic micro-cutting action of material removal leads to the formation of thin ribbon-like chips even when machining brittle solids such as glass, ferrites, or MgO. Figure 1 shows such a chip formed during the polishing of soda-lime glass. Machined surfaces of ceramics such as Ni-Zn ferrite, soda-lime glass, Si_3N_4, and zirconia created by this plastic mechanism of removal do contain residual stresses of the order of 20–50 MPa (Chandrasekar *et al* 1991). Furthermore, dislocation etch-pitting experiments on ground and polished blocks of single crystal MgO show the presence of an intensely deformed surface layer with high dislocation densities. Both these observations reinforce the hypothesis of a plastic micro-cutting mechanism of material removal as well as the formation of microindentations and plastic scratches on machined ceramic surfaces. The plastic material removal mechanism in nominally brittle solids is no doubt a consequence of the large hydrostatic stresses generated in a small volume underneath the ceramic surface by the abrasive particles.

When larger loads are transmitted through the abrasives, a transition is seen to occur in
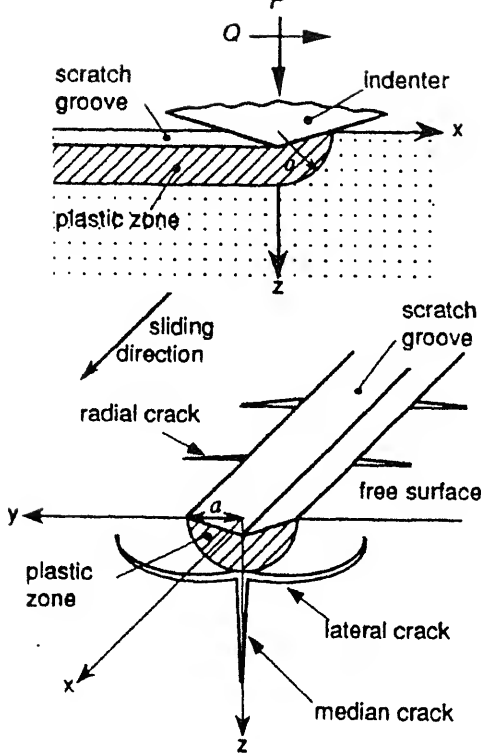
Figure 2. Schematic of sliding indentation and resulting fracture.

brittle fracture. The most common modes of material removal by brittle fracture are due to lateral cracks breaking open onto the surface, grain pull-out, or a crushing-type of material removal. Again, these conclusions have been reached from microscopic observations of machined surfaces and wear particles.

In order to understand better these mechanisms, and the driving forces behind them, it is worthwhile to analyse a reduced model for the machining of ceramics. This is provided by the sliding indentation process which entails a single abrasive particle sliding across the ceramic work-piece. This configuration has been examined extensively in the literature (Broese van Groenou *et al* 1979; Swain 1979; Evans and Marshall 1981; Cheng and Finnie 1990; Ahn *et al* 1993). A schematic of this experiment and the resulting crack pattern are shown in figure 2. As summarized by Ahn *et al* (1993), the experimental results delineate different force regimes in which the various crack patterns illustrated in figure 2 occur. For a Vickers indenter sliding on soda-lime glass, the results can be summarized as follows. At normal loads less than $\sim$ 0.05 N, no cracking is observed but a groove is formed indicating localized plastic deformation. The formation of such grooves in some cases is a consequence of material removal by a plastic-cutting mechanism, which produces chips similar to those shown in figure 1. In other cases, the groove is merely a sliding indent, i.e. the material within the groove is displaced mostly to the sides of the groove. In this latter instance there is no material removal but just a plastic indentation. In the force range of 0.05–0.8 N, a median crack is observed perpendicular to the surface. The depth of the median crack increases with normal force. In the range 0.8–3 N median cracking is

accompanied by lateral cracking parallel to the surface. At loads in the higher portion of this range, lateral cracks break through to the surface causing material removal. At still higher loads, say 3–6 N there is considerable crushing of particles in the scratch groove, along with a median crack.

A stress analysis that approximates the localized inelastic deformation using a sliding blister field has been completed by Ahn *et al* (1993). This was obtained by extending the blister field model for static indentation proposed by Yoffe (1982). In this model the strength of the blister field is evaluated as a function of the indentation load by measuring the volume of the scratch groove (figure 2). The value of the tensile stress in the uncracked solid at the location of the different systems is reported by Ahn *et al* (1993). The stress that would initiate a lateral crack is found to be small for low loads and equals the median crack driving stress near loads at which lateral cracks are typically observed to occur. That is, the sliding blister model accurately predicts the experimentally observed critical load for lateral crack formation in soda-lime glass. To date, a complete stress analysis of the sliding indentation of brittle materials does not exist.

It must be observed here that the sliding indentation forces in the results summarized above are in the range of forces applied to abrasives during lapping and polishing. The lapping force calculations described earlier yield average particle forces in the range 0.03–0.6 N for the lapping and polishing of glasses, ferrites, and Al_2O_3 under typical conditions. The smaller forces correspond to abrasive particle sizes of ~ 1 μm while the higher forces correspond to particle sizes of ~ 63 μm. The smaller of the force values are in the range where plastic material removal by cutting has been observed both during polishing and sliding indentation. The higher particle forces are well in the range of conditions for which lateral cracking is a dominant mechanism of material removal in sliding indentation experiments.

The calculation of volumetric wear as a function of process variables in abrasive machining processes has not been carried out to any significant extent. Evans & Marshall (1981) used fracture mechanics to obtain a formula for wear volume as a function of applied force and ceramic material properties, based on a mechanism of lateral cracking to describe material removal. However, their formulae for predicting the onset of lateral cracking as well as volumetric removal rates have not been validated by experiment, e.g. Larsen-Basse (1994). The analytical estimation of such wear rates during the grinding, lapping, and polishing of ceramics, based on the different wear mechanisms that are operative, is a problem worthy of detailed study.

4. Discussion

This short review of recent work on abrasive machining of ceramics has highlighted the two main mechanisms of wear associated with material removal: (1) brittle fracture with lateral cracks intersecting the machined surface, when the load applied by an abrasive particle is high, and (2) ductile micro-cutting with chip formation taking place as in single point machining of metals. The evidence for these mechanisms has come from microscopic observations of machined surfaces, chips and wear particles; from a consideration of the forces and pressures applied by the individual abrasive particles in machining, and from

the nature of deformation occurring in ceramic surfaces under sliding microindenters subjected to loads similar to those acting on abrasive particles. Other observations pertaining to residual stresses, magnetic property changes, microcracking and strength of machined ceramics provide further support for these mechanisms. For example, lapped and finely polished surfaces of crystalline ceramics typically have residual compressive stresses and high dislocation densities in shallow surface layers which support a mechanism of material removal by plastic cutting, and material displacement by indentation. Furthermore, such surfaces show little evidence of microcracking and there is very little strength degradation and strength anisotropy in lapped and finely polished ceramics. In contrast, coarsely polished or ground surfaces of ceramics show considerable microcracking in the surface layers as well as strength anisotropy and strength degradation. This is consistent with a mechanism of material removal and wear taking place by brittle fracture.

In conclusion, the cost effective machining of ceramics necessitates an understanding of the deformation processes and stress fields in ceramics produced by a sliding indenter under conditions of pressure and temperature prevalent in machining. By controlling and exploiting the transition between wear particle formation by ductile cutting and by brittle fracture, it should be possible to increase material removal rates and significantly improve the quality of the machined surface. The use of active chemical reagents to enhance or decrease material removal through tribochemical reactions at the interface between the abrasives and the ceramic workpiece should also be beneficial in this regard.

References

Ahn Y, Farris T N, Chandrasekar S 1993 Elastic stress fields caused by sliding microindentation of brittle materials. In *Machining of advanced materials*. NIST SP 847 (ed.) S Jahanmir, pp 71–81

Braza J F, Cheng H S, Fine M E, Gangopadhyay A K, Keer L, Worden R E 1989 Mechanical failure mechanisms in ceramic sliding and rolling contacts. *Tribol. Trans.* 32: 1–8

Broese van Groenou A, Maan N, Veldkamp J B D 1979 Single-point scratches as a basis for understanding grinding and lapping. In *The science of ceramic machining and surface finishing* (eds) B J Hockey, R W Rice, NBS SP 562, vol. 2, pp 43–60

Chandrasekar S, Farris T N, Shaw M C, Bhushan B 1991 Surface finishing processes for magnetic recording head ceramics. *ASME Adv. Infomation Storage Systems* 1 (1): 353–373

Chauhan R, Ahn Y, Chandrasekar S, Farris T N 1993 Role of indentation fracture in free abrasive machining of ceramics. *Wear* 162: 246–257

Cheng W, Finnie I 1990 A mechanism for sub-surface median crack initiation in glass during indenting and scribing. *J. Mater. Sci.* 25: 575–579

Evans A G, Marshall D B 1981 Wear mechanisms in ceramics. In *Fundamentals of friction and wear of materials* (ed.) D A Rigney (ASM) pp 439–452

Hebbar R R, Chandrasekar S, Farris T N 1992 Ceramic grinding temperatures. *J. Am. Ceram.*

cker S A, Farris T N, Chandrasekar S 1994 Technique for measuring dynamic grinding contact stiffness and effective wheel modulus. *J. Tribol.* (submitted)

1anmir S, Dong X 1994 Wear mechanisms of aluminum oxide ceramics. In *Friction and wear of ceramics* (Marcel Dekker) pp 15–49

1anmir S, Ives L K, Ruff A W, Peterson M B 1992 Ceramic machining: Assessment of current practice and research needs in the United States. Technical report, NIST SP 834

rsen Basse J 1994 Abrasive wear of ceramics. In *Friction and wear of ceramics* (ed.) S Jahanmir (Marcel Dekker) pp 99–115

L, Farris T N, Chandrasekar S 1992 Sliding microindentation wear particles: Spheres in grinding swarf. In *From the cradle to the grave* (eds) D Dowson, C M Taylor, T H C Childs, M Godet, G Dalmaz (Elsevier) pp 257–263

arshall D B, Evans A G, Khuri-Yakub B T, Tien J W, Kino G S 1983 The nature of machining damage in brittle materials. *Proc. R. Soc.* A385: 461–475

okes R J 1972 The effect of surface finishing on mechanical and other physical properties of ceramics. In *The science of ceramic machining and surface finishing* (eds) R W Rice, S J Schneider, NBS SP 348, pp 343–353

bramanian K 1988 Precision finishing of ceramic components with diamond abrasives. *Am. Ceram. Soc. Bull.* 67: 1026–1029

vain M V 1979 Microfracture about scratches in brittle solids. *Proc. R. Soc.* A366: 575–597

offe E H 1982 Elastic stress fields caused by indenting brittle materials. *Philos. Mag.* A46: 617–628

Optimization

Foreword

It is with great pleasure that we present to the Indian engineering community this special issue of *Sādhanā* on optimization. Optimization pervades most, if not all, engineering endeavours, cutting across interdisciplinary barriers. It can manifest in cost minimization, minimization of energy loss in a power system or delays in a communication network, and so on. It can be 'static' as in linear programming or 'dynamic' as in control. In all its multiple avatars, it remains a forever young and vibrant discipline that never ceases to surprise us with newer, more powerful techniques and conceptual breakthroughs.

In its pedagogical treatments, the word 'optimization' has somehow come to be associated with its operations research aspects. But that is only one facet of it: Error minimization in statistical signal processing is as much of an optimization problem as, say, a transportation problem. We have taken this broader view of the subject and the wide spectrum of articles will reflect this. Needless to say, the choice is limited both by what can be achieved within the confines of a single issue and, even more so, by the biases inherent in our own backgrounds. Thus it represents optimization issues in operations research, electrical engineering and artificial intelligence. Within this context, we have tried to get a very representative picture of where the subject is and where it is headed. Our hope is that we have succeeded "not wholly or in full measure, but very substantially". If so, a lion's share of the credit goes to our contributors. We take this opportunity to express our heartfelt gratitude to them. We also thank our referees for their valuable time and effort. We thank Ms. Shashikala, Ms. Nair and their colleagues at *Sādhanā* for the effort they have put into the making of this issue. Finally, we record our special thanks to Prof. N Viswanadham, Editor, *Sādhanā* , who initiated the process and encouraged us all the way.

August 1997

V S BORKAR
VIJAY CHANDRU
S SATHIYA KEERTHI
Guest Editors

timal adaptive control for a class of stochastic systems

ARUNABHA BAGCHI[1] and HAN-FU CHEN[2]

[1]Department of Applied Mathematics, University of Twente, PO Box 217, 7500
AE Enschede, The Netherlands
[2]Institute of Systems Science, Chinese Academy of Sciences, Beijing 100 080,
People's Republic of China
e-mail: a.bagchi@math.utwente.nl

Abstract. We study linear-quadratic adaptive tracking problems for a spe-
cial class of stochastic systems expressed in the state-space form. This is a
long-standing problem in the control of aircraft flying through atmospheric tur-
bulence. Using an ELS-based algorithm and introducing dither in the control
law we show that the resulting control achieves optimal cost in the limit, while
simultaneously the unknown parameters converge to their true values.

Keywords. Stochastic adaptive control; tracking problem; ELS-based
estimator.

Introduction

re is an enormous amount of literature on stochastic adaptive control starting with the
neering work of Åström & Wittenmark (1973) on the self-tuning regulator. Most of the
earch in this area, however, concentrated on ARMAX models (Kumar 1985). Parallel
his was the somewhat unrelated development of the design of adaptive flight control
tems, starting with the thesis of Illiff (1973). Most researchers in this area start with
dynamical model of aircraft in flight and, consequently, formulate their problem in the
e-space form. Unfortunately, the literature on stochastic adaptive control for systems
tate-space form is rather limited. Kumar (1983) made a thorough analysis on the prob-
1 of controlling an unknown linear-Gaussian system with quadratic criterion but had to
trict himself to the case of complete observation of the system states. We study here
problem of controlling a linear system with incomplete and inaccurate observation
the system states so that a quadratic tracking criterion is minimized in the situation
en the matrix multiplying the control term in the state equation is unknown. We do not
ume that the observation noise is Gaussian, but do restrict ourselves to the situation
h no state noise. This problem arises naturally in controlling the flight of an aircraft in
ospheric turbulence where the objective is to minimize the normal acceleration or gust

and pilot comfort. It corresponds to our problem when the so-called control derivatives of the aircraft are unknown.

2.   Problem formulation

Consider the following discrete-time dynamical system

$$x_{k+1} = Ax_k + Bu_k, \tag{1}$$
$$y_k = Cx_k + w_k, \tag{2}$$

where x_k and y_k, for fixed k, are $\mathbb{R}^n$- and $\mathbb{R}^m$-valued state and observation vectors respectively; u_k is an $\mathbb{R}^p$-valued control vector, $\{w_k\}$ is a noise sequence to be specified below, the matrices A and C are known, while the matrix B is unknown. Our objective is to minimize the tracking criterion

$$J = \lim_{N \to \infty} \sup \frac{1}{N} J_N(u), \tag{3}$$

where

$$J_N(u) \triangleq \sum_{k=0}^{N-1} [(x_k - x_k^*)' Q_1 (x_k - x_k^*) + u_k' Q_2 u_k], \tag{4}$$

with $Q_1 \geq 0$, $Q_2 > 0$ and $\{x_k^*\}$ a prescribed sequence of desired path.

We follow the self-tuning approach which is based on the certainty equivalence principle. Thus we first determine the optimal control law assuming that B is known, and then replace B by an appropriate on-line estimator. Recursive estimator for B may be obtained by using extended least squares (ELS), maximum likelihood (ML), stochastic approximation (SA) or Kalman filter methods. In the specific problem considered here, the Kalman filter method yields readily a recursive estimator for B (Balakrishnan 1987). The estimator also has been proved (Balakrishnan 1987) to converge to the true value in the mean-square sense. When we close the control loop the analysis becomes more complicated. Asymptotic optimality of the resulting control law has not been established yet. One difficulty is that the estimator loses the interpretation of being the conditional espectation when the system operates in closed loop. We propose in the next section an ELS-based method to estimate B.

The rest of this section is devoted to determining the optimal control law when B is known. For this, we make the following assumptions:

A1. $\alpha^{-1}(e^{i\lambda}) + \alpha^{-1}(e^{-i\lambda}) - 1 > 0 \; \forall \lambda \in [0, 2\pi)$, where $\alpha(z) \triangleq \det(I - zA) \equiv 1 + a_1 z + \cdots + a_n z^n$, $z \in \mathbb{C}$. This is the *strict positive real* (SPR) condition. Note that this condition implies $\alpha(z) \neq 0 \; \forall |z| \leq 1$ (see Chen & Guo 1991, corollary 4.1).

A2. (A, C) is observable.

A3. (A, B) is controllable and (A, D) is observable, where D is any matrix satisfying $D'D = Q_1$.

A4. $\{w_k, \mathcal{F}_k\}$ is a martingale difference sequence (mds) with

$$\lim_{k \to \infty} \sup E[\|w_k\|^\beta | J_{k-1}] < \infty \quad \text{a.s. for some } \beta \geq 2.$$

e calculations that follow are well-known (see Bagchi 1993, for example). They are luded only for completeness. By A3, there is a unique solution to the following algebraic ccati equation in the class of positive definite matrices.

$$S = A'SA - A'SB(Q_2 + B'SB)^{-1}B'SA + Q_1 \tag{5}$$

I the matrix,

$$F \triangleq A - B(Q_2 + B'SB)^{-1}B'SA, \tag{6}$$

isymptotically stable. Define

$$L \triangleq -(Q_2 + B'SB)^{-1}B'SA, \tag{7}$$

$$b_k \triangleq -\sum_{j=0}^{\infty}(F^j)'Q_1 x_{k+j}^* = F'b_{k+1} - Q_1 x_k^*, \tag{8}$$

$$d_k \triangleq -(Q_2 + B'SB)^{-1}B'b_{k+1}. \tag{9}$$

ing (5), we have,

$$x_N'Sx_N - x_0'Sx_0 = \sum_{k=0}^{N-1}(x_{k+1}'Sx_{k+1} - x_k'Sx_k)$$

$$= \sum_{k=0}^{N-1}(Ax_k + Bu_k)'S(Ax_k + Bu_k)$$

$$- \sum_{k=0}^{N-1} x_k'[A'SA - A'SB(Q_2 + B'SB)^{-1}B'SA + Q_1]x_k,$$

that

$$\sum_{k=0}^{N-1}(x_k'Q_1 x_k + u_k'Q_2 u_k)$$

$$= x_0'Sx_0 - x_N'Sx_N + \sum_{k=0}^{N-1}(u_k'B'SBu_k + 2x_k'A'SBu_k$$

$$+ x_k'L'(Q_2 + B'SB)Lx_k + u_k'Q_2 u_k) \tag{10}$$

milarly, we obtain

$$b_N'x_N - b_0'x_0 = \sum_{k=0}^{N-1}(b_{k+1}'x_{k+1} - b_k'x_k)$$

$$= \sum_{k=0}^{N-1}[b_{k+1}'(Ax_k + Bu_k) - b_k'x_k]. \tag{11}$$

$$J_N(u) = \sum_{k=0}^{N-1} (x_k' Q_1 x_k - 2x_k^{*'} Q_1 x_k + x_k^{*'} Q_1 x_k^* + u_k' Q_2 u_k)$$

$$= x_0' S x_0 - x_N' S x_N + \sum_{k=0}^{N-1} (u_k' B' S B u_k + 2x_k' A' S B u_k$$

$$+ x_k' L'(Q_2 + B'SB)L x_k + u_k' Q_2 u_k) + 2b_0' x_0 - 2b_N' x_N$$

$$+ 2\sum_{k=0}^{N-1} [b_{k+1}'(A x_k + B u_k) - b_k' x_k] + \sum_{k=0}^{N-1} (x_k^{*'} Q_1 x_k^* - 2x_k^{*'} Q_1 x_k). \tag{12}$$

From (8) we get

$$b_k' x_k + x_k^{*'} Q_1 x_k = b_{k+1}' F x_k = b_{k+1}' A x_k + b_{k+1}' B L x_k. \tag{13}$$

Putting this in (12) yields

$$J_N(u) = x_0' S x_0 - x_N' S x_N + 2_0' x_0 - 2b_N' x_N$$

$$+ \sum_{k=0}^{N-1} (u_k' B' S B u_k + 2x_k' A' S B u_k + u_k' Q_2 u_k$$

$$+ x_k' L'(Q_2 + B'SB)L x_k + 2b_{k+1}' B u_k - 2b_{k+1}' B L x_k + x_k^{*'} Q_1 x_k^*)$$

$$= x_0' S x_0 - x_N' S x_N + 2b_0' x_0 - 2b_N' x_N + \sum_{k=0}^{N-1} (x_k^{*'} Q_1 x_k^* - d_k' B' b_{k+1})$$

$$+ \sum_{k=0}^{N-1} (u_k - L x_k - d_k)'(Q_2 + B'SB)(u_k - L x_k - d_k). \tag{14}$$

Let $\mathcal{U}$ be the class of admissible controls which will be specified below. The point to note at this moment is that, whatever class $\mathcal{U}$ we choose for admissible controls,

$$\inf_{u \in \mathcal{U}} \cdot \lim_{N \to \infty} \sup \frac{1}{N} J_N(u) \geq \lim_{N \to \infty} \sup \frac{1}{N} \sum_{k=0}^{N-1} [x_k^{*'} Q_1 x_k^* - d_k' B' b_{k+1}]. \tag{15}$$

Thus, if in the class of desired control laws we are able to choose a control for which $\lim_{N \to \infty} \sup(1/N) J_N(u)$ equals the right hand side of (15), then this will automatically yield the optimal control desired.

3. Recursive estimator for B

Let

$$C \text{ adj } (I - zA) \equiv C + C_1 z + \cdots + C_{n-1} z^{n-1}, \tag{16}$$

here "adj" stands for the adjoint of a matrix and C_i are $m \times n$ matrices, $i = 1, \ldots, n-1$.
t

$$\theta' = [CB \vdots C_1 B \vdots \cdots \vdots C_{n-1} B], \tag{17}$$

d

$$\phi'_k = [u'_k \vdots u'_{k-1} \vdots \cdots \vdots u'_{k-n+1}], \tag{18}$$

here CB and $C_i B$ are $m \times p$ matrices.
Let us define

$$\Sigma \triangleq \begin{bmatrix} C \\ C_1 \\ \vdots \\ C_{n-1} \end{bmatrix} \text{ and } \Pi = \begin{bmatrix} CB \\ C_1 B \\ \vdots \\ C_{n-1} B \end{bmatrix}. \tag{19}$$

mma 1. *With the notations above,*
$$B = (\Sigma' \Sigma)^{-1} \Sigma' \Pi.$$

oof. All we have to show is that Σ is of full row-rank so that $\Sigma' \Sigma$ is invertible. Suppose
at

$$\Sigma x = 0, \quad x \in \mathbb{R}^n.$$

en

$$C \text{ adj } (I - zA)x = 0,$$

nich implies that

$$C(I - zA)^{-1} x = 0.$$

en for $|z|$ sufficiently small,

$$C(I + zA + z^2 A^2 + \cdots)x \equiv 0.$$

om this we conclude that

$$Cx = 0, \quad CAx = 0, \ldots, CA^n x = 0.$$

nce (A, C) is observable, we must have $x = 0$, establishing that Σ is of full row
nk.   □

We are now in a position to propose a recursive algorithm for estimating B. We first
opose the following scheme to estimate θ recursively:

$$\theta_{k+1} = \theta_k + \gamma_k P_k \phi_k [\alpha(z) y_{k+1} - \theta'_k \phi_k - (\alpha(z) - 1) \hat{w}_{k+1}], \tag{20}$$

$$P_{k+1} = P_k - \gamma_k P_k \phi_k \phi'_k P_k, \tag{21}$$

$$\gamma_k = (1 + \phi'_k P_k \phi_k)^{-1}, \tag{22}$$

$$\hat{w}_{k+1} = \alpha(z) y_{k+1} - \theta'_{k+1} \phi_k - (\alpha(z) - 1) \hat{w}_{k+1}, \tag{23}$$

here z now stands for the unit delay operator and θ_0, P_0, $\hat{w}_0$ are chosen arbitrarily.

Let us write $\hat\theta_k$ in the block matrix form

$$\theta_k' = [\theta_{k1} \vdots \cdots \vdots \theta_{kn}]$$

where θ_{ki} are $m \times p$ matrices, $i = 1, \ldots, n$, and set

$$\Pi_k = \begin{bmatrix} \theta_{k1} \\ \vdots \\ \theta_{kn} \end{bmatrix} \tag{24}$$

We propose the following recursive estimator for B:

$$B_k = (\Sigma'\Sigma)^{-1}\Sigma'\Pi_k. \tag{25}$$

Theorem 1. *Assume that conditions* A1, A2 *and* A4 *hold. Then for any* $\mathcal{F}_k$-*measurable control* u_k,

$$\|B - B_{k+1}\|^2 = O\left(\frac{\log\lambda_{\max}(k)(\log\log\lambda_{\max}(k))^{\Delta(\beta-2)}}{\lambda_{\min}(k)}\right) \tag{26}$$

where $\lambda_{\max}(k)$ *and* $\lambda_{\min}(k)$ *denote the maximum and minimum eigenvalues of* P_{k+1}^{-1}, *respectively,* β *is as defined in* A4 *and*

$$\Delta(\beta - 2) = \begin{cases} 0 & \text{if } \beta > 2 \\ c > 1, & \text{but otherwise arbitrary, if } \beta = 2. \end{cases}$$

Proof. Using lemma 1, we have

$$B - B_k = (\Sigma'\Sigma)^{-1}\Sigma'(\Pi - \Pi_k),$$

so that it is sufficient for us to show that

$$\|\theta - \theta_{k+1}\|^2 = O\left(\frac{\log\lambda_{\max}(k)(\log\log\lambda_{\max}(k))^{\Delta(\beta-2)}}{\lambda_{\min}(k)}\right). \tag{27}$$

Note that, with z denoting the unit delay operator, we may rewrite (1) and (2) symbolically as follows:

$$x_k = (I - zA)^{-1}Bu_{k-1},$$
$$y_k = C(I - zA)^{-1}Bu_{k-1} + w_k.$$

Therefore, we get

$$\alpha(z)y_k = C \text{ adj}\,(I - zA)Bu_{k-1} + \alpha(z)w_k. \tag{28}$$

Using (17) and (18), we may express (28) as

$$\alpha(z)y_{k+1} = \theta'\phi_k + \alpha(z)w_{k+1}. \tag{29}$$

Let us denote

$$\xi_{k+1} \stackrel{\Delta}{=} \hat{w}_{k+1} - w_{k+1}.$$

Using (23) and (29), we have

$$\xi_{k+1} = \theta'\phi_k + \alpha(z)w_{k+1} - \theta'_{k+1}\phi_k - (\alpha(z) - 1)\hat{w}_{k+1} - w_{k+1}$$

$$= \tilde{\theta}'_{k+1}\phi_k - (\alpha(z) - 1)\xi_{k+1},$$

or, equivalently,

$$\alpha(z)\xi_{k+1} = \tilde{\theta}'_{k+1}\phi_k, \tag{30}$$

where $\tilde{\theta}_{k+1} \triangleq \theta - \theta_{k+1}$. Using (20) and (23) we have

$$\hat{w}_{k+1} = \xi_{k+1} + w_{k+1} = \alpha(z)y_{k+1} - (\alpha(z) - 1)\hat{w}_{k+1}$$

$$- (\theta_k + \gamma_k P_k \phi_k (\alpha(z)y_{k+1} - \theta'_k\phi_k - (\alpha(z) - 1)\hat{w}_{k+1})')'\phi_k$$

$$= (1 - \gamma_k \phi'_k P_k \phi_k)(\alpha(z)y_{k+1} - (\alpha(z) - 1)\hat{w}_{k+1} - \theta'_k\phi_k)$$

$$= \gamma_k(\alpha(z)y_{k+1} - (\alpha(z) - 1)\hat{w}_{k+1} - \theta'_k\phi_k).$$

Therefore, (20) may be rewritten as

$$\theta_{k+1} = \theta_k + P_k\phi_k(\xi_{k+1} + w_{k+1})', \tag{31}$$

or, equivalently,

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - P_k\phi_k(\xi_{k+1} + w_{k+1})'. \tag{32}$$

The result now follows from theorem 2 of Chen & Guo (1991) by identifying η_k in that theorem with,

$$\alpha(z)y_{k+1} - (\alpha(z) - 1)\hat{w}_{k+1} - \theta'_k\phi_k - w_{k+1},$$

which is easily seen to be $\mathcal{F}_k$-measurable. □

4. Consistent estimator for B

The previous theorem shows that the estimation error of B_k depends upon the behavior of P_k^{-1}. In general, we do not know whether B_k converges to the true B or not. To ensure strong consistency of the estimator of unknown parameters and achieve optimality of the control law at the same time is a very difficult problem. Direct certainty-equivalence based adaptive contol law cannot achieve this goal in the linear-quadratic problem (Polderman 1989). In stochastic adaptive control literature the idea of diminishing dither to the control law has been introduced for this purpose (Caines & Lafortune 1984; Chen 1984), which will be used here.

Let $\{v_k\}$ be a sequence of $\mathbb{R}^p$-valued random vectors which is independent of $\{w_k\}$, with $Ev_k = 0$, $Ev_kv'_k = I$ and $\|v_k\| \leq$ constant a.s. (almost surely). Define

$$u^d_k \triangleq \frac{v_k}{k^{\epsilon/2}}, \quad \epsilon \in \left[0, \frac{1}{2n}\right). \tag{33}$$

Without loss of generality, we may assume that $\mathcal{F}_k \equiv \sigma\{w_i, v_i, 0 \le i \le k\}$. Set

$$\mathcal{F}'_k \equiv \sigma\{w_i, 0 \le i \le k, \quad v_j, \quad 0 \le j \le k-1\}.$$

Let u_k^s be any $\mathcal{F}'_k$-measurable control law at time k, obtained possibly by some certainty equivalence principle. We apply the diminishingly excited version u_k of u_k^s to the system,

$$u_k = u_k^s + u_k^d. \tag{34}$$

Theorem 2. *If* A1, A2, A4 *hold, and if*

$$\sum_{i=1}^{k} \|u_i^s\|^2 = O(k^{1+\delta}), \quad \delta \in \left[0, \frac{1-2\epsilon n}{1+2n}\right), \quad a.s. \tag{35}$$

Then

$$\|B - B_k\|^2 = O\left(\frac{\log k (\log\log k)^{\Delta(\beta-2)}}{k^\alpha}\right),$$

$$\alpha \in \left(\frac{1}{2}(1+\delta), 1 - n(\epsilon + \delta)\right]. \tag{36}$$

Proof. We first note that the interval for α is not empty, because

$$n(\epsilon + \delta) + \frac{\delta}{2} < n\left(\epsilon + \frac{1-2\epsilon n}{1+2n}\right) + \frac{1-2\epsilon n}{2(1+2n)} = \frac{1}{2}.$$

It is easy to establish (see equation (6.72) in Chen & Guo 1991) that

$$\frac{1-\epsilon}{k^{1-\epsilon}} \sum_{i=1}^{k} u_k^d u_k^{d'} \to I \text{ as } n \to \infty. \tag{37}$$

Then

$$\lambda_{\max}(k) = O\left(\sum_{i=1}^{k} \|u_i\|^2\right)$$

$$= O\left(\sum_{i=1}^{k} (\|u_i^s\|^2 + \|u_i^d\|^2)\right) = O(k^{1+\delta}). \tag{38}$$

Using (26) and (38), we can establish (36) provided that

$$\lim_{k\to\infty} \inf k^{-\alpha} \lambda_{\min}\left(\sum_{i=0}^{k} \phi_i \phi'_i\right) \ne 0, \tag{39}$$

where $\lambda_{\min}(X)$ denotes the minimum eigenvalue of a matrix X. Suppose that (39) does not hold. Then there exists a sequence $\{\zeta_{k_\ell}\}$ with

$$\zeta_{k_\ell} = [\rho_{k_\ell}^{(1)'} : \cdots : \rho_{k_\ell}^{(n)'}]', \quad \|\zeta_{k_\ell}\| = 1,$$

being p-dimensional vectors, such that

$$\lim_{\ell \to \infty} k_\ell^{-\alpha} \sum_{i=1}^{k_\ell} (\zeta'_{k_\ell} \phi_i)^2 = 0,$$

equivalently,

$$\lim_{\ell \to \infty} k_\ell^{-\alpha} \sum_{i=1}^{k_\ell} (\rho_{k_\ell}^{(1)'} u_i + \cdots + \rho_{k_\ell}^{(n)'} u_{i-n+1})^2 = 0. \tag{40}$$

theorem 2.8 of Chen & Guo (1991),

$$\left\| \sum_{i=1}^{k} u_{i-j}^s u_i^{d'} \right\| = O(k^{(1+\delta/2)} (\log k)^{(1/2)+\eta}), \quad \forall j \geq 0, \quad \forall \eta > 0,$$

$$\left\| \sum_{i=1}^{k} u_{i-j} u_i^{d'} \right\| = O(k^{(1+\delta/2)} (\log k)^{(1/2)+\eta}), \quad \forall j > 0.$$

nce $\alpha > (1+\delta)/2$, we then have,

$$k_\ell^{-\alpha} \left\{ \rho_{k_\ell}^{(1)'} \sum_{i=1}^{k_\ell} u_i^s u_i^d \rho_{k_\ell}^{(1)} + \sum_{j=2}^{n} \rho_{k_\ell}^{(j)'} \sum_{i=1}^{k_\ell} u_{i-j+1} u_i^{d'} \rho_{k_\ell}^{(1)} \right\} \to 0, \quad k \to \infty,$$

hich, using (40), leads to

$$k_\ell^{-\alpha} \sum_{i=1}^{k_\ell} (\rho_{k_\ell}^{(1)'} u_i^d)^2 \to 0, \quad k \to \infty, \tag{41}$$

d

$$k_\ell^{-\alpha} \sum_{i=1}^{k_\ell} (\rho_{k_\ell}^{(1)'} u_i^s + \rho_{k_\ell}^{(2)'} u_{i-1} + \cdots + \rho_{k_\ell}^{(n)'} u_{i-n+1})^2 \to 0, \quad k \to \infty. \tag{42}$$

om (37) and (41) it follows that

$$\| \rho_{k_\ell}^{(1)} \|^2 = o(k_\ell^{-(1-\epsilon-\alpha)}), \tag{43}$$

hich, together with (35), implies that

$$k_\ell^{-(1+\delta)+1-\epsilon-\alpha} \sum_{i=1}^{k_\ell} (\rho_{k_\ell}^{(1)'} u_i^s)^2 \to 0, \quad k \to \infty. \tag{44}$$

quations (44) and (42) imply that

$$k_\ell^{-\alpha-(\epsilon+\delta)} \sum_{i=1}^{k_\ell} (\rho_{k_\ell}^{(2)'} u_{i-1} + \cdots + \rho_{k_\ell}^{(n)'} u_{i-n+1})^2 \to 0, \quad k \to \infty. \tag{45}$$

omparing (45) with (40) yields

$$\| \rho_{k_\ell}^{(i)} \|^2 = o(k_\ell^{-(1-\epsilon-\alpha-(i-1)(\epsilon+\delta))}), \quad i = 1, \ldots, n. \tag{46}$$

Noticing that $\alpha \leq 1 - n(\epsilon + \delta)$, we find that for every i, $1 \leq i \leq n$,

$$\alpha \leq 1 - i(\epsilon + \delta) = 1 - \epsilon - \delta - (i-1)(\epsilon + \delta) < 1 - \epsilon - (i-1)(\epsilon + \delta).$$

Therefore, (46) implies that

$$\rho_{k\ell}^{(i)} \longrightarrow 0, \quad \ell \to \infty, \quad i = 1, \ldots, n.$$

This contradicts our assumption that $\|\rho_{k\ell}\| = 1$ for all ℓ. This, in turn, contradicts (40) establishing our result.   □

5. Optimal adaptive control

Let us now go back to the adaptive control problem posed in § 2. It is clear from (14) that, if $\{x_k\}$ was completely observed and $\{B_k\}$ was known, the optimal control would be given by

$$u_k = Lx_k + d_k. \tag{47}$$

We use the ELS-based estimator B_k for B and define the certainty-equivalence control u^0 by

$$u_k^0 = L_k \hat{x}_k + \hat{d}_k, \tag{48}$$

where

$$\hat{x}_{k+1} = A\hat{x}_k + B_k u_k^0, \tag{49}$$

$\hat{x}_0$ arbitrary,

$$L_k = -(Q_2 + B_k' S_k B_k)^{-1} B_k' S_k A, \tag{50}$$

$$S_k = A' S_{k-1} A - A' S_{k-1} B_k (Q_2 + B_k' S_{k-1} B_k)^{-1} B_k' S_{k-1} A + Q_1, \tag{51}$$

$S_0 \geq 0$, otherwise arbitrary,

$$\hat{d}_k = -(Q_2 + B_k' S_k B_k)^{-1} B_k \hat{b}_{k+1}, \tag{52}$$

$$\hat{b}_k = -\sum_{j=0}^{k} F_{k-1}^j Q_1 x_{k+j}^*, \tag{53}$$

$$F_k = A - B_k L_k. \tag{54}$$

We now define stopping times $\{\sigma_k\}$ and $\{\tau_k\}$ as follows: Set $\tau_1 = 1$. Let

$$\sigma_k = \sup \left\{ t > \tau_k \,\Big|\, \sum_{i=\tau_k}^{j-1} \|u_i^0\|^2 \leq (j-1)^{1+\delta} + \|u_{\tau_k}^0\|^2, \ \forall j \in (\tau_k, t) \right\}, \tag{55}$$

$$\tau_{k+1} = \inf \left\{ t > \sigma_k \,\Big|\, \sum_{i=\tau_k}^{\sigma_k - 1} \|u_i^0\|^2 \leq \frac{t^{1+\delta}}{2^k} \bigwedge \sum_{i=\sigma_k}^{t} \|u_i^0\|^2 \leq \frac{t^{1+\delta}}{2^k} \right\}. \tag{56}$$

The desired control law u_k^s is defined by

$$u_k^s = \begin{cases} u_k^0, & \text{if } k\epsilon(\tau_\ell, \sigma_\ell) \text{ for some } \ell, \\ 0, & \text{otherwise,} \end{cases} \tag{57}$$

finally, the adaptive control law we are after is given by

$$u_k^* = u_k^s + u_k^d. \tag{58}$$

$\mathcal{U}$ denote the class of admissible controls defined by

$$\mathcal{U} = \{u \,|\, u_k \text{ is } \mathcal{F}_k - \text{measurable such that the}$$
$$\text{resulting state satisfies } \|x_k\|^2 = o(k) \ a.s.\}$$

eorem 3. *Assume that* A1 − A4 *hold. Then* $\{u_k^*\}\epsilon\mathcal{U}$,

$$\lim_{N\to\infty} \sup \frac{1}{N} J_N(u^*) = \lim_{N\to\infty} \sup \frac{1}{N} \sum_{k=0}^{N-1} [x_k^{*'} Q x_k^* - d_k' B' b_{k+1}] \tag{59}$$

d

$$\|B - B_k\|^2 = O\left(\frac{\log k (\log\log k)^{\Delta(\beta-2)}}{k^\alpha}\right), \quad \alpha \in \left(\frac{1}{2}, 1 - n\epsilon\right] \tag{60}$$

oof. We first show consistency of B_k. If $\tau_\ell < \infty$ and $\sigma_\ell = \infty$ for some ℓ, then $\sum_{i=1}^k \|u_i^s\|^2 = O(k^{1+\delta})$ and the strong consistency of B_k follows from theorem 2. If $< \infty$ and $\tau_{\ell+1} = \infty$ for some ℓ, then $u_i^s = 0 \ \forall i \ge \ell$ and, again by theorem 2, B_k is ongly consistent.
Consider now that $\tau_\ell < \infty, \sigma_\ell < \infty$ for all ℓ. Then

$$\sup_{\tau_\ell \le k < \tau_{\ell+1}} \frac{1}{k^{1+\delta}} \sum_{i=1}^k \|u_i^s\|^2 = \sup_{\tau_\ell \le k < \sigma_\ell} \frac{1}{k^{1+\delta}} \sum_{i=1}^k \|u_i^0\|^2$$

$$= \sup_{\tau_\ell \le k < \sigma_\ell} \frac{1}{k^{1+\delta}} \left[\sum_{i=\tau_1}^{\sigma_1-1} \|u_i^0\|^2 + \sum_{i=\sigma_1}^{\tau_2} \|u_i^0\|^2 + \cdots \right.$$
$$\left. + \sum_{i=\sigma_{\ell-1}}^{\tau_\ell} \|u_i^0\|^2 + \sum_{i=\tau_\ell+1}^{k} \|u_i^0\|^2 \right]$$

$$\le \sup_{\tau_\ell \le k < \sigma_\ell} \frac{1}{k^{1+\delta}} \left[2 \sum_{i=1}^{\ell-1} \frac{\tau_{i+1}^{1+\delta}}{2^i} + \sum_{i=\tau_\ell+1}^{k} \|u^0\|^2 \right]$$

$$\le 2 + \sup_{\tau_\ell \le k < \sigma_\ell} \frac{1}{k^{1+\delta}} \sum_{i=\tau_\ell}^{k} \|u_i^0\|^2$$

$$\le 2 + \sup_{\tau_\ell \le k < \sigma_\ell} \frac{1}{k^{1+\delta}} (k^{1+\delta} + \|u_{\tau_\ell}^0\|^2)$$

$$\le 3 + \frac{\|u_{\tau_\ell}^0\|^2}{\tau_\ell^{1+\delta}} \le 3 + \frac{1}{2^{\ell-1}} < 4.$$

Therefore, $\sum_{i=1}^{k} \|u_i^s\|^2 = O(k^{1+\delta})$ and theorem 2 again leads to the strong consistency of B_k.

By theorem 3.4 and remark 3.3 of Chen & Guo (1991),

$$S_k \to S, k \to \infty,$$

from which it also follows that

$$F_k \to F, \quad \|b_k - \hat{b}_k\| \to 0 \quad \text{and} \quad \|d_k - \hat{d}_k\| \to 0, \quad k \to \infty.$$

Since F is a stable matrix and $F_k \to F$, there exists a $\rho \in (0, 1)$ and a $C > 0$ such that

$$\|F_k F_{k-1} \cdots F_1\| \le C\rho^k \quad \forall k. \tag{61}$$

We now show that there exists an ℓ_0 such that

$$\tau_{\ell_0} < \infty \quad \text{and} \quad \sigma_{\ell_0} = \infty. \tag{62}$$

If $\sigma_\ell < \infty$ and $\tau_{\ell+1} = \infty$ for some ℓ, then $u_i^* = u_i^d \ \forall i \ge \sigma_\ell$ and from (49) it follows that $\{\hat{x}_k\}$ is bounded. Hence $\{u_k^0\}$ is bounded and by definition (56) $\tau_{\ell+1}$ must be finite. This means that $\sigma_\ell < \infty$ and $\tau_{\ell+1} = \infty$ cannot happen for any ℓ.

If $\sigma_\ell < \infty$ and $\tau_\ell < \infty$ for all ℓ, then for $i \in [\tau_\ell, \sigma_{\ell-1}]$,

$$\hat{x}_{i+1} = A\hat{x}_i + B_i(L_i\hat{x}_i + \hat{d}_i + u_i^d)$$
$$= F_i\hat{x}_i + B_i(\hat{d}_i + u_i^d),$$

so that

$$\hat{x}_{i+1} = F_i F_{i-1} \cdots F_{\tau_\ell}\hat{x}_{\tau_\ell} + \sum_{j=1}^{i-\tau_\ell} F_i F_{i-1} \cdots F_{i-j+1}B_{i-j}(\hat{d}_{i-j} + u_{i-j}^d)$$
$$+ B_i(\hat{d}_i + u_i^d).$$

Using this and (61) we have

$$\|\hat{x}_{i+1}\| \le C\rho^{i-\tau_\ell+1}\|\hat{x}_{\tau_\ell}\| + C_1 \sum_{j=0}^{i-\tau_\ell} \rho^j$$
$$= C\rho^{i-\tau_\ell+1}\|\hat{x}_{\tau_\ell}\| + C_2,$$

so that

$$\|\hat{x}_{i+1}\|^2 \le 2C\rho^{2(i-\tau_\ell+1)}\|\hat{x}_{\tau_\ell}\|^2 + 2C_2^2.$$

Therefore,

$$\sum_{i=\tau_\ell-1}^{\sigma_\ell-1} \|\hat{x}_{i+1}\|^2 \le C_3\|\hat{x}_{\tau_\ell}\|^2 + C_4\sigma_\ell,$$

or

$$\sum_{i=\tau_\ell}^{\sigma_\ell} \|\hat{x}_i\|^2 \le C_3\|\hat{x}_{\tau_\ell}\|^2 + C_4\sigma_\ell = o(\sigma_\ell^{1+\delta}) \text{ as } \ell \to \infty.$$

s result, along with (48), gives us

$$\sum_{i=\tau_\ell}^{\sigma_\ell} \|u_i^o\|^2 = o(\sigma_\ell^{1+\delta}) \quad \text{as } \ell \to \infty.$$

s means that for sufficiently large ℓ,

$$\sum_{i=\tau_\ell}^{\sigma_\ell} \|u_i^0\|^2 < \sigma_\ell^{1+\delta},$$

ich contradicts the definition of σ_ℓ, since by (55)

$$\sum_{i=\tau_\ell}^{\sigma_\ell} \|u_i^0\|^2 > \sigma_k^{1+\delta} + \|u_{\tau_\ell}^0\|^2.$$

us we cannot have $\tau_\ell < \infty$ and $\sigma_\ell < \infty$ for all ℓ.
Consequently (62) holds and

$$u_k^* = u_k^0 + u_k^d,$$

$\hat{x}_{k+1} = F_k \hat{x}_k + B_k \hat{d}_k, \quad \forall k \geq \tau_{\ell_0}.$
By (61) and the boundedness of $B_k \hat{d}_k$ it is clear that $\{\hat{x}_k\}$ is also bounded and so is $\{u_k^0\}$,
that

$$\sum_{i=1}^{k} \|u_i^s\|^2 = O(k).$$

is and theorem 2 imply (60).
To prove optimality, notice that

$$x_{k+1} - \hat{x}_{k+1} = A(x_k - \hat{x}_k) + (B - B_k)u_k^*.$$

the boundedness of $\{u_k^*\}$, the stability of A and the fact that $B - B_k \to 0$ we find that

$$x_k - \hat{x}_k \longrightarrow 0, k \to \infty. \tag{63}$$

is, along with the boundedness of $\{\hat{x}_k\}$, implies that $\{x_k\}$ is also bounded. Therefore,
$_k^*\} \in \mathcal{U}$. Using (63) and the facts that $L_k - L \to 0$, and $\|d_k - \hat{d}_k\| \to 0$, we conclude
at

$$\frac{1}{N} \sum_{k=0}^{N-1} (u_k^* - Lx_k - d_k)'(Q_2 + B'SB)(u_k^* - Lx_k - d_k) \longrightarrow 0, \quad k \to \infty.$$

ombining this with (14) establishes (59). $\square$

Conclusion

e solved a class of stochastic adaptive control problems in the state space form which
ise in controlling aircraft flying in gusty conditions. The important, although difficult,
tensions which should be further looked into involve the state noise case and/or when
e parameters A and C also contain unknown elements. The approach presented in this
aper does not directly go over to this most general situation, but may possibly be used
ere in combination with some other techniques.

References

Åström K J, Wittenmark B 1973 On self-tuning regulators. *Automatica* 9: 185–199

Bagchi A 1993 *Optimal control of stochastic systems* (London: Prentice Hall International)

Balakrishnan A V 1987 *Kalman filtering theory* (New York: Optimization Software)

Caines P, Lafortune S 1984 Adaptive control with recursive identification for stochastic linear systems. *IEEE Trans. Autom. Control* AC-29: 312–321

Chen H F 1984 Recursive system identification and adaptive control by use of the modified least squares algorithm. *SIAM J. Control Optimization* 22: 758–776

Chen H F, Guo L 1991 *Identification and stochastic adaptive control* (Boston: Birkhäuser)

Illiff K W 1973 *Identification and stochastic control with application to flight control in turbulence*. Ph D thesis, University of California, Los Angeles

Kumar P R 1983 Optimal adaptive control of linear quadratic Gaussian systems. *SIAM J. Control Optimization* 21: 163–178

Kumar P R 1985 A survey of some results in stochastic control. *SIAM J. Control Optimization* 23: 329–380

Polderman J W 1989 *Adaptive control and identification: Conflict or conflux?* CWI Tract 67, (Amsterdam: Centre for Mathematics & Computer Science)

Wang G H 1993 *A stochastic adaptive control application to flight systems*. Ph D thesis, University of California, Los Angeles

ntrol of a 2-DOF manipulator with a flexible forearm

KOH TUCK LYE, H KRISHNAN* and C L TEO

Department of Mechanical and Production Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
e-mail: mpehk@nus.sg (H Krishnan)

Abstract.   In this paper we present experimental results on the position and vibration control of the end-effector of a 2-DOF parallelogram manipulator with a flexible forearm. A dynamic model of the manipulator is first obtained. Control strategies are implemented to control the manipulator. The first control strategy uses the computed torque method based on a reduced-order dynamic model of the manipulator which is obtained by assuming that all the links are rigid. This method is referred to as the reduced-order computed torque (ROCT) method. Experimental results demonstrate that such a strategy is not good for vibration control of the end-effector of the manipulator. The second control strategy is a state feedback control law designed based on a local linearization of the nonlinear dynamic model. Experimental results show that this control strategy achieves good vibration control of the end-effector of the manipulator. However, since the strategy is based on local linearization, it is valid only in a neighbourhood of the operating point. A hybrid controller that uses the ROCT method for the initial large movement of the manipulator is then implemented. Based on a switching rule, the controller is switched to the state feedback controller based on the linearized model when the manipulator is sufficiently close to the equilibrium state. Experimental results are reported and the successful performance of the controller in dampening out end-point vibrations is demonstrated.

Keywords.   Two-DOF flexible manipulator; reduced-order computed torque control; vibration control; hybrid controller.

Introduction

e increasing industrial need for higher productivity requires robots that move accurately rapidly. In order to achieve this, it is desirable to have robots with light structures offer high-speed performance and low energy consumption. One way of designing a ot with a light structure is to use slender links. Modelling and control of flexible-link

thor for correspondence

manipulators has been a subject of active research in recent years. Flexible-link manipulators have the advantages of being lightweight and have a larger pay-load to weight ratio. They are also easy to fabricate and their design is simple. However the fact that the links are flexible introduces vibrations as the robot arm is in motion. Vibration suppression and position control in flexible-link robot arms is therefore very important. A number of researchers have studied the control design for vibration suppression and position control of a flexible robot arm (see Craig 1981, Book 1984, Cannon & Schmitz 1984, Hastings & Book 1986, Sakawa & Matsuno 1986, Matsuno & Fukushima 1987, Siciliano & Book 1988, Wang & Vidyasagar 1991, Siciliano et al 1992, De Luca & Siciliano 1993, Aoustin et al 1994, Gross & Tomizuka 1994, Khorrami et al 1994, Moudgal et al 1994, Vandergrift et al 1994, Zhu et al 1994, Banavar & Dominic 1995, Lin & Yih 1996, and references therein). Successful experimental results have also been reported for single-link and two-link flexible robot arms (Cannon & Schmitz 1984; Hastings & Book 1986; Sakawa & Matsuno 1986; Matsuno & Fukushima 1987; Aoustin et al 1994; Gross & Tomizuka 1994; Khorrami et al 1994; Moudgal et al 1994; Banavar & Dominic 1994). Modelling of flexible robot arms has been carried out in many papers (see Book 1984, for e.g.). It has been pointed out (Siciliano & Book 1988) that, if the links are stiff, one can use singular perturbation methods for the control design for vibration suppression and position control in n-link flexible manipulators. However the control law given by Siciliano & Book (1988) requires measurements of the link deflection rates which is not easily measurable. The control law is modified as an output feedback strategy by Siciliano et al (1992) and only requires the joint position measurements, joint velocities, and the link deflection, which are obtained easily, say, from strain gauge measurements. An alternative but similar control design method is also proposed by Vandergrift et al (1994) for vibration suppression and position control of n-link flexible manipulators. A method of control design based on a dynamic output feedback compensator is proposed by Craig (1981), which avoids the need for the velocity of the deflection. Sakawa & Matsuno (1986) derived a dynamic model for a 2-DOF parallelogram manipulator and designed a feedback control system for it. This was followed up by Matsuno & Fukushima (1987) who obtained several satisfactory results for point-to-point position control.

The objective of this paper is to report experimental results on the position and vibration control of the end-effector of a 2-DOF parallelogram manipulator with a flexible forearm. A dynamic model is first derived. Control strategies are implemented on the experimental facility. The first control strategy uses the computed torque method based on a reduced-order dynamic model of the manipulator which is obtained by assuming that all the manipulator links are rigid. This method is referred to as the reduced-order computed torque (ROCT) method. The second control strategy is a state feedback control law designed based on a local linearization of the nonlinear dynamic model of the system. Finally, the third control strategy uses the ROCT control for the initial large movement of the manipulator. The control law is then switched to the state feedback control based on a switching rule. This method is referred to as the hybrid control method.

The paper is organised as follows. Section 2 describes the experimental setup. This is followed by the derivation of the dynamic model of the system in § 3. Using this dynamic model, three different control strategies are developed. Section 4 presents the experimental
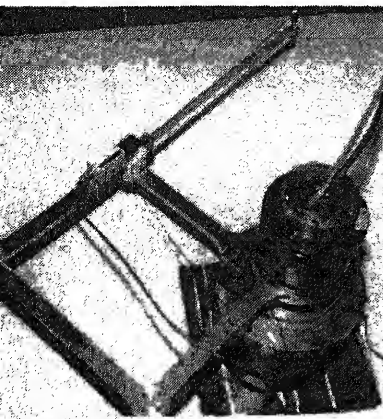
Figure 1. The manipulator with a flexible forearm.

lts of the ROCT method. In § 5, the state feedback control law and hybrid control
are discussed and the experimental results are given. The conclusions are presented
6.

Experimental setup

section provides a brief description of the experimental setup. The objective of this
er is to present experimental results on the position and vibration control of the end-
ctor of a 2-DOF manipulator with a flexible forearm. Figure 1 shows a picture of the
ipulator in the experimental facility.

he manipulator is a 2-DOF parallelogram design similar to the one developed by
ls & Lokhorst (1993). The last link (forearm) of the robot is flexible. It operates in
horizontal plane and is driven by 2 Dynaserv direct-drive brushless motors so that
problem of friction and backlash are minimized. The characteristics of the motors
d in the experimental robot are shown in table 1. Note that the motors have high
ormance, high torque output which prevents saturation, and high precision encoder
lution. The servo drivers for the motors can be preset to operate in 3 modes of
trol, namely, position control mode, speed control mode and torque/current control
le. In this experiment we preset the operation of the servo drivers to the torque/current
trol mode. Thus a voltage between -8 V to 8V applied to the servo drivers allows
motor to produce a torque between negative maximum torque and positive maximum

le 1. Motor characteristics.

ure	Unit	Motor model	
		DM1200A	DM1045B
imum output torque	Nm	200	45
d speed	rev/s	1.0	2.0
oder resolution	p/rev	1024000	655360
ght	kg	29	9.5
or inertia	kgm^2	0.167	0.019
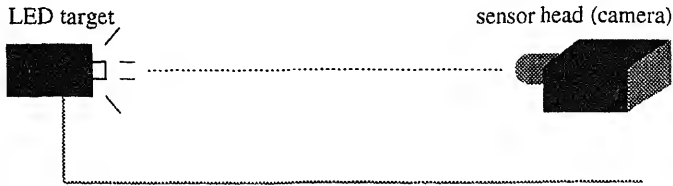
Figure 2. Schematic diagram of an active type PSD.

torque, and the voltage-to-torque relationship is linear. Thus direct torque control is achieved.

In this experiment, measurement of the vibration at the end-effector of the robot is essential. The Hamamatsu C2399-00 position-sensitive detector (PSD) is used for this purpose. This is an active type, opto-electric position sensing unit which measures the position of a single-point of light spot focused on the sensor head. It uses an infrared LED mounted as a target at the tip of the flexible link and the movement of the LED is measured by a sensor camera which is mounted as shown in figure 1. Figure 2 shows the schematic diagram of the active type PSD used in the experiment. The advantage of the C2399-00 PSD is that it uses a non-discrete, silicon PSD for two-dimensional position sensing. The non-discrete nature of the PSD allows for high accuracy measurement of position. The LED is pulse-modulated and optically filtered in the sensor head with a built-in background-light cancellation circuit, making accurate measurements possible even in bright locations. Since it does not require scanning, it provides good speed of response. The advantages of high accuracy measurement and good speed of response make it very suitable for vibration measurement. The rated position resolution of the C2399-00 PSD is 1/5000 and its sampling rate is 312.5 Hz. The sensor head of the PSD is a camera that uses a C-mount lens. The choice of a suitable lens is based on the following 2 factors: (1) the required angle of view which is determined by the dimensions of the scene, and (2) the minimum illumination level of the area to be monitored.
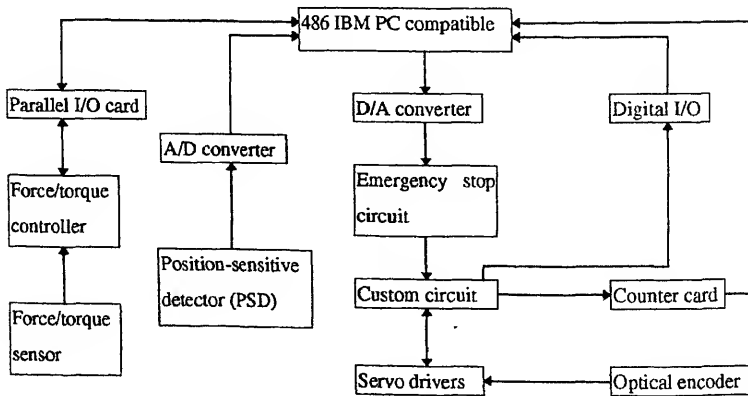
Figure 3 shows the controller architecture. A quadrature encoder input card along with the motor encoders are used for measurement of link angular displacements. The link angular velocities are then obtained by numerical differentiation. Inputs to this card are conditioned by a 4-stage digital filter. For the experiment, the encoder signals from the 2 motors are input to the card, which provides the digital counts to the computer. The host computer is an IBM PC-compatible machine with a 486DX33 microprocessor with 8MB RAM. A program written in C-language is used to control the manipulator. A menu-driven control panel allows the user to input control parameters such as desired end position of manipulator, sampling rate, and gain constants. At each sampling time, the computer reads the digital input data, performs servo calculations and writes digital output data.

The D/A converter for the digital computer operates in bipolar mode and has a 14-bit resolution and conversion time less than $42\,\mu s$. The voltage output range of the D/A converter is $-8V\ +8V$ and its outputs are connected to the servo drivers via an emergency stop circuit and a custom circuit. These are used for controlling the voltages to the two motors. The A/D converters provide analog-to-digital conversion for the analog signal from the PSD. The A/D converter has a 12-bit resolution and best conversion time of $10\,\mu s$. The accuracy and linearity of the converter are both $+/-1$ bit. Besides the D/A and A/D converters, the digital I/O card is also used as an interface to the computer. They are used to read the status of motors 1 and 2 respectively. An interface to the force/torque sensor is also available as shown in figure 3. This is used for conducting contact force control experiments using the manipulator.

3. Dynamic model of the manipulator

3.1 *Manipulator characteristics and parameters*

Figure 4 shows a schematic diagram of the manipulator with a flexible forearm. It comprises 4 rigid links labelled from links 1 to 4. Link 4 has an extension l_s that is used to clamp flexible link 5. All the links are made of aluminum alloy 5083. A position-sensitive detector
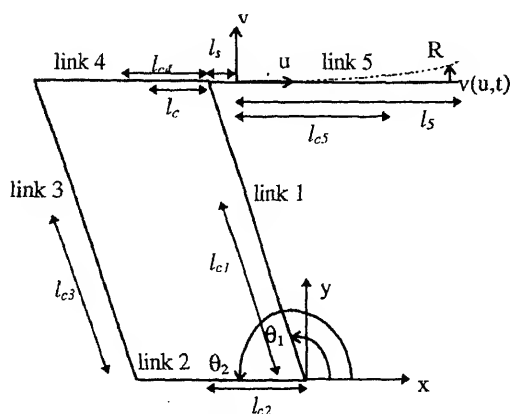
Figure 4. Schematic diagram of the manipulator.

	Camera assembly	LED assembly
Length (m)	$l_c = 0.0645$	$l_s = 0.0525, l_5 = 0.4800$
Mass (kg)	$m_c = 0.209$	$m_L = 0.026$

(PSD) is mounted on the manipulator. The PSD consists of a LED assembly which is mounted at the tip of link 5 and a camera assembly mounted at a distance l_c from the joint of links 4 and 1.

Table 2 shows the parameters of the camera assembly and the LED assembly. Table 3 shows the parameters for each link of the parallelogram manipulator.

3.2 Free transverse vibration of Bernoulli–Euler beam

In order to derive the dynamic equation of the manipulator, which comprises a flexible forearm labelled as link 5 in figure 4, it is essential to know its vibrational characteristics. Assuming link 5 is a Bernoulli–Euler beam, its vibration is governed by the Bernoulli–Euler beam equation given by

$$\frac{\partial^2}{\partial u^2}\left(EI\frac{\partial^2 v}{\partial u^2}\right) + \rho\frac{\partial^2 v}{\partial t^2} = 0, \tag{1}$$

where $0 \le u \le l_5$, ρ is the linear density of the beam, E is Young's modulus and I is the cross-sectional area moment of inertia of link 5. The values of ρ and I for link 5 are $0.244\,\text{kg/m}$ and $6.75 \times 10^{-11}\,\text{m}^4$ respectively. The assumed-modes method is used in our analysis. The deflection of link 5 is given in separable form as

$$v(u, t) = \sum_{i=1}^{\infty} \phi_i(u)q_i(t), \tag{2}$$

where $\phi_i(u)$ is the ith natural mode eigenfunctions, and $q_i(t)$ is the time-dependent generalized coordinate. The above equations are solved by approximating the natural modes of flexible link 5 by the natural modes of a uniform clamped-free beam with boundary conditions given by (Craig 1981)

Table 3. Parameters of each individual link.

	Link 1	Link 2	Link 3	Link 4	Link 5
Width (mm)	30	30	30	30	3
Height (mm)	15	15	15	15	30
Length (m)	$l_1 = 0.40$	$l_2 = 0.35$	$l_3 = 0.40$	$l_4 = 0.35$	$l_5 = 0.48$
Centre of mass (m)	$l_{c1} = 0.058$	$l_{c2} = 0.090$	$l_{c3} = 0.195$	$l_{c4} = 0.173$	$l_{c5} = 0.293$
Mass (kg)	$m_1 = 2.905$	$m_2 = 1.505$	$m_3 = 0.877$	$m_4 = 0.858$	$m_5 = 0.117$
Moment of inertia (kg/m²)	$I_1 = 0.079$	$I_2 = 0.031$	$I_3 = 0.023$	$I_4 = 0.019$	$I_5 = 2.24 \times 10^{-3}$

Table 4. Natural frequencies and related constants of link 5.

de i	Mode 1	Mode 2	Mode 3	Mode 4
	1.87510	4.69409	7.85476	10.99554
	0.7341	1.0185	0.9992	1.0000
$_5$)	2.00	−2.00	2.00	−2.00
rad/s)	69.64	436.49	1540.89	3019.55
Hz)	11.08	69.47	245.24	480.58
$m_L = 0$)				

$$\nu|_{u=0} = 0, \quad \frac{\partial \nu}{\partial u}\bigg|_{u=0} = 0, \quad \frac{\partial^2 \nu}{\partial u^2}\bigg|_{u=l_5} = 0, \quad \frac{\partial^3 \nu}{\partial u^3}\bigg|_{u=l_5} = 0. \tag{3}$$

e natural mode shape eigenfunctions and the natural frequencies are given by (Craig
31)

$$\phi_i(u) = [\cosh \beta_i u - \cos \beta_i u - \sigma_i(\sinh \beta_i u - \sin \beta_i u)], \tag{4}$$

$$\omega_i = \beta_i^2 (EI/\rho)^{0.5}, \tag{5}$$

ere

$$1 + \cos \beta_i l_5 \cosh \beta_i l_5 = 0, \tag{6}$$

$$\sigma_i = \frac{\cos \beta_i l_5 + \cosh \beta_i l_5}{\sin \beta_i l_5 + \sinh \beta_i l_5}, \tag{7}$$

$i = 1, \cdots \infty$.

Table 4 shows the natural frequencies and related constants for the first 4 modes of
k 5. Note that the natural frequencies listed in table 4 are for link 5 without point mass
ts tip.

Manipulator dynamics

e method used in the dynamic modelling of the system is the Lagrange–Euler formula-
n. In this method, the computation of the kinetic and potential energy of the system is
integral part of the formulation. As the detailed derivation of the dynamic equations is
ry long and tedious, only the relevant equations are presented here. For more details on
derivation, please refer to appendix A.

The full physical system comprises links 1 through 5, 2 direct drive brushless motors
d a camera-LED assembly mounted on the arm. The dynamic model of the system is
en by (please refer to appendix A)

$$D(Q)\ddot{Q} + GQ + H(Q, \dot{Q}) + \mu(\dot{\theta}_1, \dot{\theta}_2) = PT_m, \tag{8}$$

$$Q = [\theta_1 \theta_2 q_1 q_2 \cdots q_\infty]^T, \quad T_m = [\tau_{m1} \tau_{m2}]^T,$$

$$D(Q) =$$

$$
\begin{bmatrix}
d_1 + J_{m1} & M_{12} & \left(\dfrac{-2\sigma_1}{\beta_1}\rho l_1 - m_L l_1 \phi_1(l_5)\right) & \left(\dfrac{-2\sigma_2}{\beta_2}\rho l_1 - m_L l_1 \phi_2(l_5)\right) & \cdots \\
 & & \times \cos(\theta_2 - \theta_1) & \times \cos(\theta_2 - \theta_1) & \\[2ex]
 & d_2 + J_{m2} + \rho l_5 \displaystyle\sum_{i=1}^{\infty} q_i^2 & & & \\
M_{21} & \qquad\qquad & \alpha_1 + m_L(l_s + l_5)\phi_1(l_5) & \alpha_2 + m_L(l_s + l_5)\phi_2(l_5) & \cdots \\[1ex]
 & + m_L \displaystyle\sum_{i=1}^{\infty} \phi_i(l_5) q_i & & & \\[1ex]
 & \displaystyle\sum_{j=1}^{\infty} \phi_j(l_5) q_j & & & \\[2ex]
\left(\dfrac{-2\sigma_1}{\beta_1}\rho l_1 - m_L l_1 \phi_1\right) & \alpha_1 + m_L(l_s + l_5)\phi_1(l_5) & \rho l_5 + m_L \phi_1(l_5)^2 & m_L \phi_1(l_5)\phi_2(l_5) & \cdots \\
\times \cos(\theta_2 - \theta_1) & & & & \\[2ex]
\left(\dfrac{-2\sigma_2}{\beta_2}\rho l_1 - m_L l_1 \phi_2\right) & \alpha_2 + m_L(l_s + l_5)\phi_2(l_5) & m_L \phi_1(l_5)\phi_2(l_5) & \rho l_5 + m_L \phi_2(l_5)^2 & \cdots \\
\times \cos(\theta_2 - \theta_1) & & & & \\[2ex]
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
$$

$$
G = \begin{bmatrix}
0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & \cdots \\
0 & 0 & \rho l_5 \omega_1^2 & 0 & \cdots \\
0 & 0 & 0 & \rho l_5 \omega_2^2 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
$$

$$
\mu(\dot{\theta}_1, \dot{\theta}_2) = \begin{bmatrix}
B_{m1}\dot{\theta}_1 + b_{m1}\,\mathrm{sgn}(\dot{\theta}_1) \\
B_{m2}\dot{\theta}_2 + b_{m2}\,\mathrm{sgn}(\dot{\theta}_2) \\
0 \\
0 \\
\vdots
\end{bmatrix}, \quad
P = \begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
0 & 0 \\
\vdots & \vdots
\end{bmatrix},
$$

$$H(Q, \dot{Q}) =$$

$$
\begin{bmatrix}
-d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_2^2 + 4\rho l_1 \dot{\theta}_2 \sin(\theta_2 - \theta_1) \displaystyle\sum_{i=1}^{\infty} \frac{\sigma_i}{\beta_i}\dot{q}_i \\[2ex]
+ 2m_L l_1 \sin(\theta_2 - \theta_1)\dot{\theta}_2 \displaystyle\sum_{i=1}^{\infty} \phi_i(l_5)\dot{q}_i \\[2ex]
+ \left(2\rho l_1 \displaystyle\sum_{i=1}^{\infty} \frac{\sigma_i}{\beta_i} q_i + m_L l_1 \displaystyle\sum_{i=1}^{\infty} \phi_i(l_5) q_i\right) \cos(\theta_2 - \theta_1)\dot{\theta}_2^2
\end{bmatrix},
$$

$$
\begin{bmatrix}
d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 + 2\rho l_5 \dot{\theta}_2 \sum_{i=1}^{\infty} q_i \dot{q}_i + 2m_L \sum_{i=1}^{\infty} \phi_i(l_5)q_i \sum_{j=1}^{\infty} \phi_j(l_5)\dot{q}_j \dot{\theta}_2 \\[2mm]
\quad - \left(2\rho l_1 \sum_{i=1}^{\infty} \frac{\sigma_i}{\beta_i} q_i + m_L l_1 \sum_{i=1}^{\infty} \phi_i(l_5)q_i \right) \cos(\theta_2 - \theta_1)\dot{\theta}_1^2 \\[2mm]
\quad - \left(2\frac{\sigma_1}{\beta_1}\rho l_1 + m_L l_1 \phi_1(l_5) \right) \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 \\[2mm]
\quad - \left(\rho l_5 q_1 + m_L \left(\sum_{i=1}^{\infty} \phi_i(l_5)q_i \right) \phi_1(l_5) \right) \dot{\theta}_2^2 \\[2mm]
\quad - \left(2\frac{\sigma_2}{\beta_2}\rho l_1 + m_L l_1 \phi_2(l_5) \right) \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 \\[2mm]
\quad - \left(\rho l_5 q_2 + m_L \left(\sum_{i=1}^{\infty} \phi_i(l_5)q_i \right) \phi_2(l_5) \right) \dot{\theta}_2^2 \\[2mm]
\vdots
\end{bmatrix},
$$

$$
\begin{aligned}
d_1 &= I_1 + m_1 l_{c1}^2 + I_3 + m_3 l_{c3}^2 + m_4 l_1^2 + m_5 l_1^2 + m_c l_1^2 + m_L l_1^2, \\
d_2 &= I_2 + m_2 l_{c2}^2 + I_4 + m_3 l_2^2 + m_4 l_{c4}^2 + \tfrac{1}{3} m_5 l_5^2 + m_5 l_s^2 \\
&\quad + m_5 l_s l_5 + m_c l_c^2 + m_L (l_s + l_5)^2, \\
d_3 &= m_3 l_2 l_{c3} - \tfrac{1}{2} m_5 l_1 l_5 + m_4 l_1 l_{c4} - m_5 l_s l_1 - m_L l_1 (l_s + l_5) + m_c l_1 l_c, \\
M_{12} &= M_{21} = d_3 \cos(\theta_2 - \theta_1) \\
&\quad + \left[2\rho l_1 \sum_{i=1}^{\infty} \frac{\sigma_i}{\beta_i} q_i + m_L l_1 \sum_{i=1}^{\infty} \phi_i(l_5)q_i \right] \sin(\theta_2 - \theta_1), \\
\alpha_i &= 2\rho \left(l_s \frac{\sigma_i}{\beta_i} + \frac{1}{\beta_i^2} \right).
\end{aligned}
$$

Note that the vector T_m denotes the vector of motor torques, J_{m_1}, J_{m_2} denote the rotor inertias of the motors which is given in table 1, B_{m_1}, B_{m_2} denote the viscous damping coefficients of the motors, and b_{m_1}, b_{m_2} denote the friction coefficients of the motors. The values for all the other parameters in the dynamic model (8) may be obtained from the data in tables 2, 3 and 4.

In (8), $D(Q)$ is the inertia matrix, $G(Q)$ is the stiffness matrix, $H(Q, \dot{Q})$ is a matrix containing nonlinear centrifugal and coriolis terms, and $\mu(\dot{\theta}_1, \dot{\theta}_2)$ is the friction and damping matrix. It should be noted that ω_i in the stiffness matrix refers to the free natural frequency of vibration of a clamped-free uniform Bernoulli–Euler beam. It does not indicate the frequency of vibration of flexible link 5. The Rayleigh's method for approximating the fundamental frequency of a continuous system is presented in § 4. This method will be used to estimate the frequency of vibration of link 5.

4. Control design based on reduced-order computed torque method

In this section, a controller is designed based on a reduced-order dynamic model of the manipulator using the computer torque control design methodology. This method is referred to as the reduced-order computed torque (ROCT) method. For control of a fully rigid manipulator arm, it is only necessary to control the rigid modes of the manipulator and the computed torque method can be used effectively for this purpose. An attempt to use the same model-based control law in a manipulator with link flexibility may result in deterioration in performance because the controller has been designed to only control the rigid modes but not the flexible modes of the manipulator. The ROCT method may be used to control a flexible-link manipulator with minimal deterioration in performance, if the discarded terms in the model-based control law have only a negligible effect on the dynamics of the system. For example, friction, viscous damping or flexibility in the system may be ignored if the system is well lubricated or is fairly rigid.

The objective of this experiment is to study the deterioration in performance, if any, of the controller designed based on the ROCT method. The physical system is described by the matrix equation given by (8). Assuming link 5 is rigid, $q_1, q_2, \ldots, q_\infty$ and $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_\infty$ are set equal to zero in (8) so that the resulting equation is the dynamic model of an equivalent 'rigid' manipulator. This equation is given by

$$
\begin{bmatrix} \tau_{m1} \\ \tau_{m2} \end{bmatrix} = \begin{bmatrix} d_1 + J_{m1} & d_3 \cos(\theta_2 - \theta_1) \\ d_3 \cos(\theta_2 - \theta_1) & d_2 + J_{m2} \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} -d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_2^2 \\ d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 \end{bmatrix}
$$
$$
+ \begin{bmatrix} B_{m1} & 0 \\ 0 & B_{m2} \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} b_{m1}\mathrm{sgn}(\dot{\theta}_1) \\ b_{m2}\mathrm{sgn}(\dot{\theta}_2) \end{bmatrix}. \tag{9}
$$

Based on the reduced-order dynamic model given by (9) and ignoring the coulomb friction and viscous damping terms which are not known exactly, a control law is obtained based on the computed torque method which is given by

$$
\begin{bmatrix} \tau_{m1} \\ \tau_{m2} \end{bmatrix} = M(\theta_1, \theta_2) \begin{bmatrix} \ddot{\theta}_{1d} + k_{p1}e_1 + k_{v1}\dot{e}_1 + k_{i1} \int_0^t e_1 \mathrm{d}t \\ \ddot{\theta}_{2d} + k_{p2}e_2 + k_{v2}\dot{e}_2 + k_{i2} \int_0^t e_2 \mathrm{d}t \end{bmatrix} + h(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)
$$
$$
\tag{10}
$$

where

$$
M(\theta_1, \theta_2) = \begin{bmatrix} d_1 + J_{m1} & d_3 \cos(\theta_2 - \theta_1) \\ d_3 \cos(\theta_2 - \theta_1) & d_2 + J_{m2} \end{bmatrix}
$$
$$
h(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2) = \begin{bmatrix} -d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_2^2 \\ d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 \end{bmatrix}
$$
$$
e_1 = \theta_{1d} - \theta_1
$$
$$
e_2 = \theta_{2d} - \theta_2
$$

The control law given by (10) is implemented to study position and vibration control of the end effector of the manipulator. Figure 5 shows the controller based on ROCT method which has been implemented on the system. Figure 6 shows the experimental results of
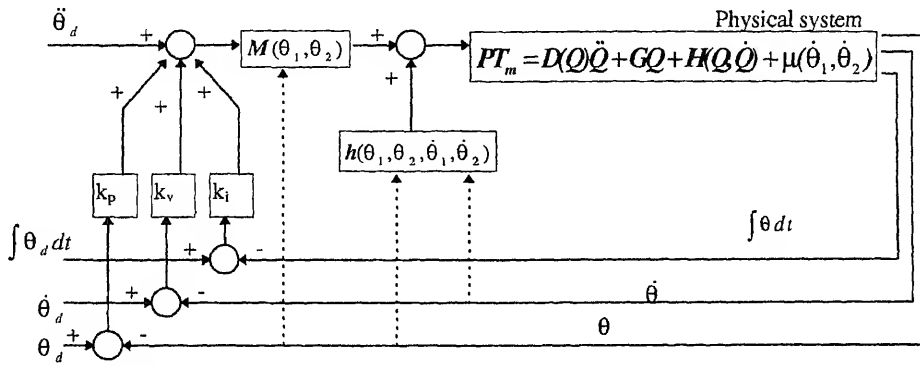
Figure 5. Controller based on reduced- order computed torque method.
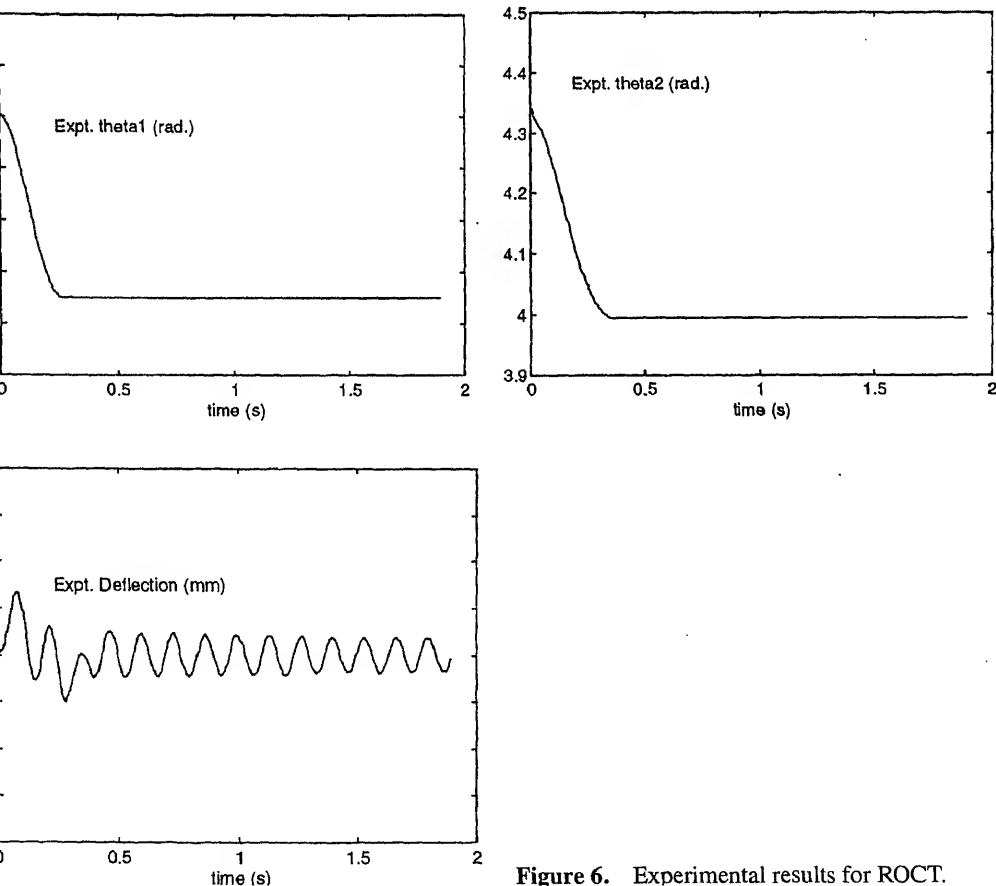
Figure 6. Experimental results for ROCT.
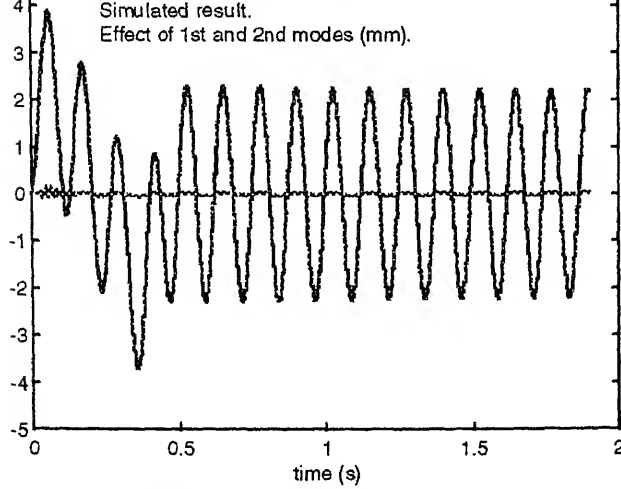
Figure 7. Relative magnitude of $\phi_1(l_5)q_1(t)$ and $\phi_2(l_5)q_2(t)$.

the ROCT method. Note that the desired state of the system is defined as $\theta_1 = 2.65$rad, $\theta_2 = 3.99$rad. The gains are set to $Kp_1 = 100$, $Kv_1 = 1$ and $Kp_2 = 50$, $Kv_2 = 4.25$.

Figure 6 shows that there is deterioration in the performance of the controller based on the ROCT method. Although the controller is able to control the rigid body modes of the manipulator, it is unable to control the vibratory modes of flexible link 5. Thus it is necessary to use a different control law for position and vibration control of the manipulator.

Rayleigh's method for approximation of the fundamental frequency of vibration is used to estimate the first fundamental frequency of vibration. Rayleigh's quotient is given by

$$\omega_i = \left\{ \int_0^{l_5} EI\phi_i''(u)^2 du \Big/ \int_0^{l_5} \rho\phi_i(u)^2 du + m_L\phi_i(l_5) \right\}^{1/2} \tag{11}$$

Since $\int_0^{l_5} EI\phi_i''(u)^2 du = \omega_i^2 \rho l_5$, $\int_0^{l_5} \rho\phi_i(u)^2 du = \rho l_5$ and all the other parameters are known, the first fundamental frequency of vibration can be estimated. The parameter values were substituted into (11) and ω_1 was found to be 8.09 Hz. From figure 6, it is observed that the natural frequency of vibration is about 7–8 Hz, which is very close to ω_1. Hence, it can be deduced that, for the experimental system, the dominant mode of vibration is mode 1. This can also be verified by checking the relative magnitude of $\phi_1(l_5)q_1(t)$ and $\phi_2(l_5)q_2(t)$. Figure 7 shows the relative magnitude of $\phi_1(l_5)q_1(t)$ and $\phi_2(l_5)q_2(t)$ obtained from simulation. From the plots in figure 7, it is obvious that $\phi_1(l_5)q_1(t)$ is large compared to $\phi_2(l_5)q_2(t)$. The deflection at the tip of link 5 can thus be approximated by

$$v(l_5, t) = \phi_1(l_5)q_1(t). \tag{12}$$

This is because mode 1 is the dominant mode of vibration and the effects of mode 2 and other higher order modes are negligible and can be discarded.

Control design based on linear state feedback method

controller based on the reduced-order computed torque (ROCT) method has been pre-
ed in the previous section. In this section a control strategy based on local linearization
ne nonlinear dynamic model of the physical system and linear state feedback control
hod is presented. Then another control strategy, referred to as the hybrid controller,
ch combines the positive characteristics of both the ROCT controller and state feedback
trol law is presented.

ecall that the matrix representation that describes the dynamics of the experimental
em is given by

$$D(Q)\ddot{Q} + GQ + H(Q, \dot{Q}) + \mu(\theta_1, \dot{\theta}_2) = PT_m. \tag{13}$$

as been deduced in § 4 that the dominant mode of vibration for the experimental sys-
is mode 1. The effect of mode 2 and other higher order modes is negligible so that
may be ignored. In general, for other systems, more than one mode may be required
give a reasonable approximation for vibration. When more than one mode is used,
all states are measurable based on measurements from the encoders and the PSD.
refore an observer must be constructed so as to estimate the unmeasurable states. In
a case, the design of the controller becomes more complex because if the system
f nth order, the observer is also of nth order for a full order state observer. The re-
ing closed loop system becomes of order $2n$; that is the complexity and size of the
olem is doubled. Furthermore, some deterioration in performance may be observed
ne dynamic response of the observer is not good. If only a 1-mode approximation is
d, it eliminates the need for an observer which is clear from (12). For the experimen-
system, a controller based on this model is implemented and satisfactory results are
ained.

Define the state vector to be $x = [\theta_1 \theta_2 q_1 \dot{\theta}_1 \dot{\theta}_2 \dot{q}]^T$ and the equilibrium state to be $x_0 = \theta_2^0 0 0 0 0]^T$, where θ_1^0 and θ_2^0 are the desired joint angles. The linearized equations
he experimental system with 1-mode approximation for the flexibilty is defined as
ows.

$$\delta\dot{x} = A\delta x + B\delta u,$$

$$\delta x = [\delta\theta_1 \ \delta\theta_2 \ \delta q_1 \ \delta\dot{\theta}_1 \ \delta\dot{\theta}_2 \ \delta\dot{q}_1]^T,$$

$$A = \begin{pmatrix} 0 & I \\ -\tilde{D}(Q)|_0^{-1}\tilde{G} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0\tilde{P} \\ \tilde{D}(Q)|_0^{-1}\tilde{P} \end{pmatrix},$$

a 3×3 null matrix, I is a 3×3 identity matrix,

$$\delta u = \delta T_m = [\tau_{m1} \tau_{m2}]^T, \quad \tilde{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^T,$$

and

$$\tilde{D}(Q)|_0 =$$

$$\begin{pmatrix} d_1 + J_{m1} & d_3 \cos(\theta_2^0 - \theta_1^0) & \left(\frac{-2\sigma_1}{\beta_1}\rho l_1 - m_L l_1 \phi_1(l_5)\right) \\ & & \times(\theta_2^0 - \theta_1^0) \\ d_3 \cos(\theta_2^0 - \theta_1^0) & d_2 + J_{m2} & \alpha_1 + m_L(l_s + l_5)\phi_1(l_5) \\ \left(\frac{-2\sigma_1}{\beta_1}\rho l_1 - m_L l_1 \phi_1(l_5)\right) & \alpha_1 + m_L(l_s + l_5)\phi_1(l_5) & \rho l_5 + m_L \phi_1(l_5)^2 \\ \times \cos(\theta_2^0 - \theta_1^0) & & \end{pmatrix} .$$

The state feedback controller for the linearized system is implemented and experimental results are obtained. The control law is given by

$$\begin{bmatrix} \tau_{m1} \\ \tau_{m2} \end{bmatrix} = -K \begin{bmatrix} \theta_1 - \theta_1^0 \\ \theta_2 - \theta_2^0 \\ q_1 \\ \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{q}_1 \end{bmatrix}.$$

The feedback gain matrix is chosen to be

$$K = \begin{bmatrix} 34.63 & -0.58 & 16.50 & 6.33 & 0.01 & 1.01 \\ 0.29 & 7.07 & -7.56 & 0.14 & 1.88 & -0.64 \end{bmatrix},$$

which gives the following closed loop eigenvalues for the system as

$$e = \begin{bmatrix} -0.92 \pm 53.07i \\ -6.30 \pm 5.43i \\ -4.00 \pm 3.75i \end{bmatrix}.$$

Figure 8 shows the experimental results of this controller, in which the desired state of the system is defined as

$$\theta_1^0 = 2.65 \text{rad}, \quad \theta_2^0 = 3.99 \text{rad}, \quad q_1^0 = 0, \quad \dot{\theta}_1^0 = 0, \quad \dot{\theta}_1^0 = 0, \quad \dot{q}_1^0 = 0.$$

Comparing figures 8 and 6, it is clear that slower response is observed for the rigid body modes using the state feedback control based on the linearized state space model in comparison to that obtained by the ROCT method. However, vibration at the tip of link 5 is suppressed after a very short time. The experimental results show the effect of the state feedback controller designed based on the linearized system obtained by a local linearization of the nonlinear dynamic model, given in (13), about an equilibrium state. When nonlinearities in the system are not severe, local linearization may be used as an approximation to the nonlinear model. However, such manipulator control is not well suited for the manipulator making a large movement. An alternative is to use gain scheduling, in which the controller gain matrix and the equilibrium state of the manipulator change with the manipulator as it moves.
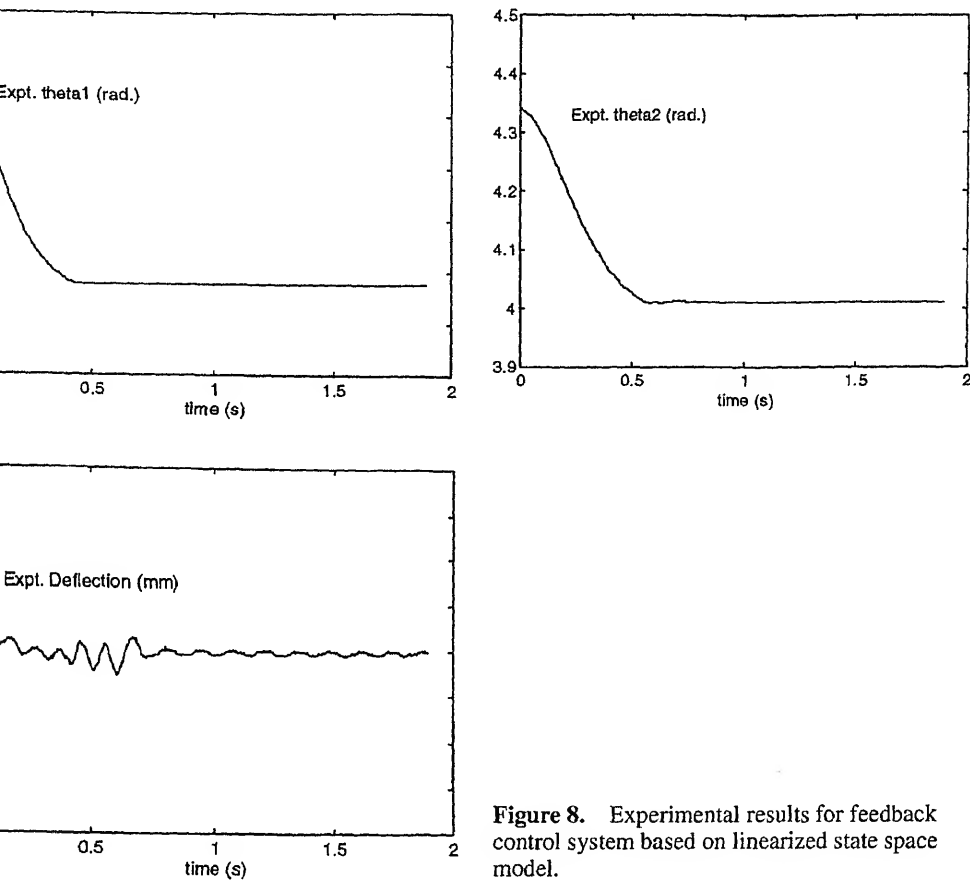
Figure 8. Experimental results for feedback control system based on linearized state space model.

hybrid controller is proposed as another alternative to control the flexible-link manipu-
The hybrid controller comprises the controller based on the reduced-order computed
e (ROCT) method for the initial large movement of the manipulator, i.e. from the
l state to a state sufficiently close to the desired state. When the manipulator is close
e desired state, the controller based on the ROCT method is switched to the linear state
back controller. A switching rule is used for such purpose. Figure 9 shows the exper-
tal results for the hybrid controller. The initial state and desired state of the system is
lentical to that defined in ROCT and state feedback control. The same gains are also
. The switching rule is to switch from ROCT controller to state feedback controller
both θ_1 and θ_2 are within $10°$ from the desired angles.

e experimental results show that there is significant improvement in the performance
e controller in terms of speed of response of the angular positions and vibration sup-
ion at the tip of the manipulator. The use of the ROCT controller for the initial large
ement of the manipulator enables fast response for the angular position. When the ma-
lator moves to a state sufficiently close to the desired state, where the local linearization
e nonlinear system is valid, the controller is switched to state feedback control. This
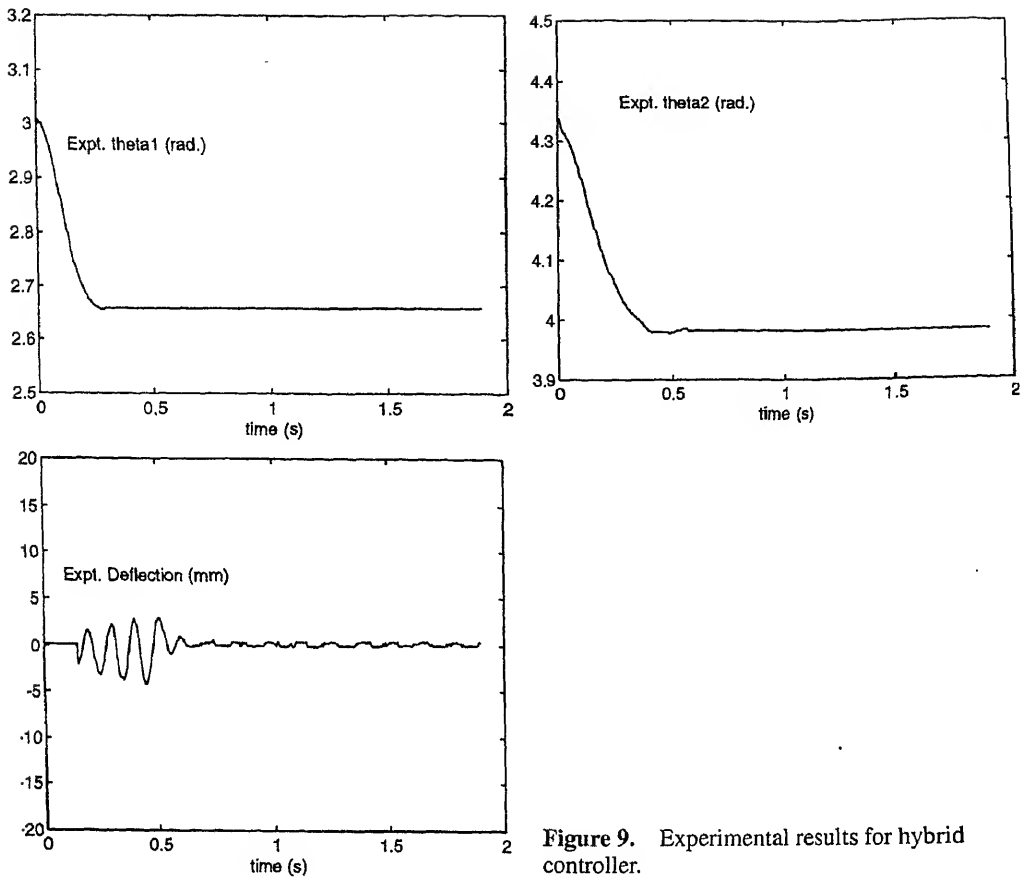
Figure 9. Experimental results for hybrid controller.

6. Conclusion

In this paper, experimental results on the position and vibration control of the end-effector of a 2-DOF manipulator with a flexible forearm have been presented. The dynamic model is obtained using the Lagrange Euler formulation and by modelling the vibration using the assumed-mode method. Control strategies are implemented to control the manipulator. The first control strategy uses the computed torque method based on a reduced-order dynamic model of the manipulator obtained under the assumption that all the links are rigid. This method has been referred to as the reduced-order computed torque (ROCT) method. Experimental results show that the ROCT controller is not suitable for position and vibration control. However, the ROCT method is good in controlling the rigid modes (angular positions) of the manipulator. The second control strategy is a state feedback control law designed based on a local linearization of the nonlinear dynamic model about an equilibrium state. Experimental results show that the performance of the state feedback control law is good and the vibration at the end-effector of the manipulator is damped out effectively. However this controller is valid only for small movement of the manipulator about the equilibrium state because linearization is done locally. The third control strategy combines the positive characteristics of each of the two controllers described above. This controller, referred to as the hybrid controller, uses the ROCT method for the initial large
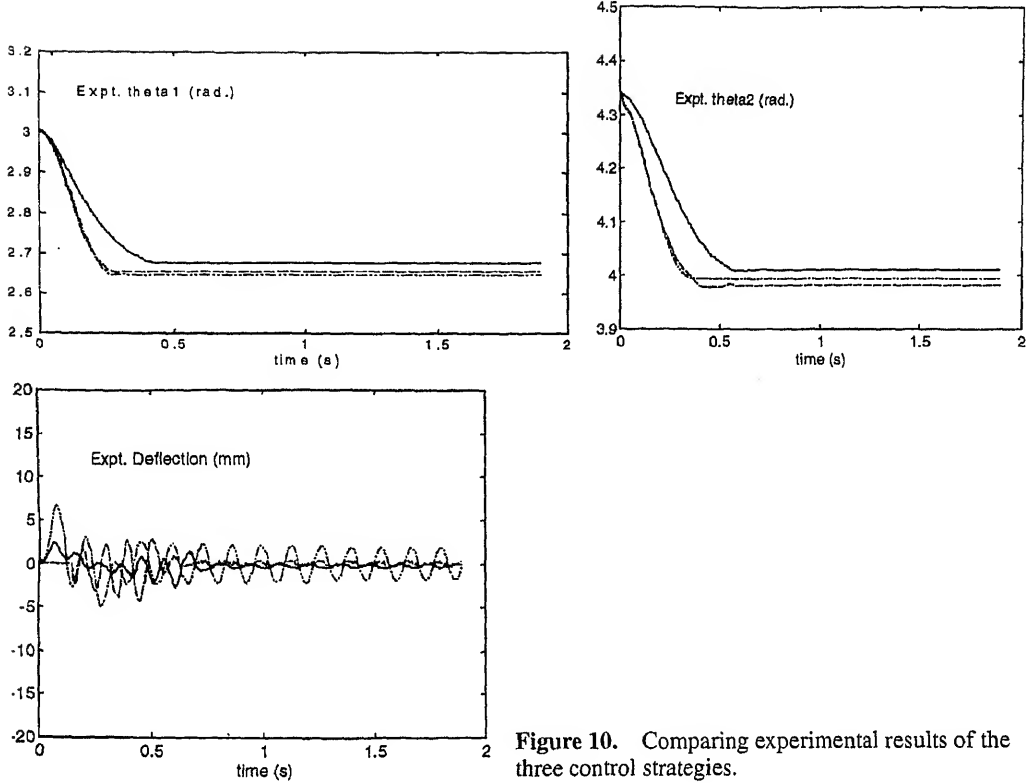
Figure 10. Comparing experimental results of the three control strategies.

movement of the manipulator. Based on a switching rule, the controller is switched to the state feedback controller when the manipulator is sufficiently close to the equilibrium state. Experimental results show good speed of response and effective vibration suppression at the end-effector of the manipulator. We mention that a one-mode approximation of the flexible modes was used in our experiments in control of the 2-DOF manipulator. If more than one mode of vibration is dominant in the system, then for vibration control one must design a state estimator for estimating the flexible modes of the manipulator based on the deflection measurements obtained using the PSD.

Appendix A.

In this section we provide the detailed derivation of the equations of motion for the 2-DOF manipulator with a flexible forearm. The full dynamic model of the system includes modelling of manipulator, modelling of point masses and modelling of the motors. The following assumptions are made on the flexible link 5 (see figure 1), in the derivation of the dynamic equations of the manipulator:

(1) Link 5 is linearly elastic and vibrates horizontally.

(2) Shear deformation and rotary inertia of link 5 are negligible.

(3) Deflection of link 5 is small and change in length of the beam is negligible.

potential energy of link 5.

Lagrange's equation is given by

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{z}_k}\right) - \frac{\partial T}{\partial z_k} + \frac{\partial U}{\partial z_k} = F_k, \quad k = 1, 2, 3, \ldots, \infty, \tag{A1}$$

where z_k is the generalized coordinate and F_k is the generalized force. The velocity of an arbitrary point R on link 5 (see figure 1) is given by

$$\tilde{v}_R = \begin{bmatrix} -l_1 \sin\theta_1 & (l_s + u)\sin\theta_2 \\ l_1 \cos\theta_1 & -(l_s + u)\cos\theta_2 \end{bmatrix}\begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} \sin\theta_2(\partial v/\partial t) + v\dot{\theta}_2 \cos\theta_2 \\ -\cos\theta_2(\partial v/\partial t) + v\dot{\theta}_2 \sin\theta_2 \end{bmatrix}.$$

The kinetic energy of links 1 to 5 are given by

$$T_1 = \tfrac{1}{2}I_1\dot{\theta}_1^2 + \tfrac{1}{2}m_1 l_{c1}^2 \dot{\theta}_1^2, \tag{A2}$$

$$T_2 = \tfrac{1}{2}I_2\dot{\theta}_2^2 + \tfrac{1}{2}m_2 l_{c2}^2 \dot{\theta}_2^2, \tag{A3}$$

$$T_3 = \tfrac{1}{2}I_3\dot{\theta}_1^2 + \tfrac{1}{2}m_3[l_{c3}^2\dot{\theta}_1^2 + l_2^2\dot{\theta}_2^2 + 2l_2 l_{c3}\dot{\theta}_1\dot{\theta}_2 \cos(\theta_2 - \theta_1)], \tag{A4}$$

$$T_4 = \tfrac{1}{2}I_4\dot{\theta}_2^2 + \tfrac{1}{2}m_4[l_1^2\dot{\theta}_1^2 + l_{c4}^2\dot{\theta}_2^2 + 2l_1 l_{c4}\dot{\theta}_1\dot{\theta}_2 \cos(\theta_2 - \theta_1)], \tag{A5}$$

$$T_5 = \frac{1}{2}\int \tilde{v}_{Rx}^2 + \tilde{v}_{Ry}^2 \, dm$$

$$= \frac{1}{2}\rho \int_0^{l_s}\left(-l_1\dot{\theta}_1 \sin\theta_1 + (l_s + u)\dot{\theta}\sin\theta_2 + \sin\theta_2\frac{\partial v}{\partial t} + v\dot{\theta}_2 \cos\theta_2\right)^2$$

$$+ \left(l_1\dot{\theta}_1 \cos\theta_1 - (l_s + u)\dot{\theta}_2 \cos\theta_2 - \cos\theta_2\frac{\partial v}{\partial t} + v\dot{\theta}_2 \sin\theta_2\right)^2 du.$$

Since link 5 is a uniform beam,

$$m_5 = \rho l_5,$$

$$I_b = \int_0^{l_s} u^2 dm = \int_0^{l_s} u^2 \rho du = \frac{1}{3}m_5 l_5^2.$$

Moreover, $v(u, t) = \sum_{i=1}^{\infty}\phi_i(u)q_i(t)$, therefore by expansion and simplification we get

$$T_5 = \frac{1}{2}m_5 l_1^2\dot{\theta}_1^2 + \frac{1}{2}[I_b + m_5 l_s^2 + m_5 l_s l_5]\dot{\theta}_2^2$$

$$- \frac{1}{2}m_5[l_1 l_5 + 2l_1 l_s]\dot{\theta}_1\dot{\theta}_2 \cos(\theta_2 - \theta_1)$$

$$+ \frac{1}{2}\rho\int_0^{l_s}\left[\sum_{i=1}^{\infty}\phi_i(u)\dot{q}_i(t)\sum_{j=1}^{\infty}\phi_j(u)\dot{q}_j(t)\right]$$

$$+ \left[\sum_{i=1}^{\infty}\phi_i(u)q_i(t)\sum_{j=1}^{\infty}\phi_j(u)q_j(t)\right]\dot{\theta}_2^2$$

$$+ 2\left[-l_1\dot{\theta}_1 \cos(\theta_2 - \theta_1)\sum_{i=1}^{\infty}\phi_i(u)\dot{q}_i(t)\right.$$

$$+ l_1 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i(u) q_i(t) \Bigg]$$

$$+ 2 \Bigg[(l_s + u) \dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(u) \dot{q}_i(t) \Bigg] du. \tag{A6}$$

In the experiment, a camera-LED assembly is mounted on the manipulator. The effect of the LED assembly, mounted at the tip of flexible link 5, is to decrease the natural frequency of vibration of link 5. For modelling accuracy, the camera and LED are modelled as point masses. The following kinematic relationships are used in the derivation.

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{bmatrix} l_1 \cos \theta_1 + l_c \cos \theta_2 \\ l_1 \sin \theta_1 + l_c \sin \theta_2 \end{bmatrix},$$

$$\begin{bmatrix} \dot{x}_c \\ \dot{y}_c \end{bmatrix} = \begin{bmatrix} -\dot{\theta}_1 l_1 \sin \theta_1 \dot{\theta}_2 l_c \sin \theta_2 \\ \dot{\theta}_1 l_1 \cos \theta_1 + \dot{\theta}_2 l_c \cos \theta_2 \end{bmatrix},$$

$$\begin{bmatrix} x_e \\ y_e \end{bmatrix} = \begin{bmatrix} l_1 \cos \theta_1 - (l_s + l_5) \cos \theta_2 + v(l_5) \sin \theta_2 \\ l_1 \sin \theta_1 - (l_s + l_5) \sin \theta_2 - v(l_5) \cos \theta_2 \end{bmatrix},$$

$$\begin{bmatrix} \dot{x}_e \\ \dot{y}_e \end{bmatrix} = \begin{bmatrix} -\dot{\theta}_1 l_1 \sin \theta_1 + \dot{\theta}_2 (l_s + l_5) \sin \theta_2 + v(l_5) \dot{\theta}_2 \cos \theta_2 + \dot{v}(l_5) \sin \theta_2 \\ \dot{\theta}_1 l_1 \cos \theta_1 - \dot{\theta}_2 (l_s + l_5) \cos \theta_2 + v(l_5) \dot{\theta}_2 \sin \theta_2 - \dot{v}(l_5) \cos \theta_2 \end{bmatrix}.$$

Note that (x_c, y_c) refers to the location of the camera and (x_e, y_e) refers to the location of the LED, which is at the tip of link 5 (see figure 1). The kinetic energy of the sensor camera is given by

$$T_c = \tfrac{1}{2} m_c (\dot{x}_c^2 + \dot{y}_c^2)$$
$$= \tfrac{1}{2} m_c [l_1^2 \dot{\theta}_1^2 + l_c^2 \dot{\theta}_2^2 + 2 l_1 l_c \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_2 - \theta_1)]. \tag{A7}$$

The kinetic energy of the LED is given by

$$T_{LED} = \tfrac{1}{2} m_L (\dot{x}_e^2 + \dot{y}_e^2).$$

This can be simplified as

$$T_{LED} = \frac{1}{2} m_L \Bigg[l_1^2 \dot{\theta}_1^2 + (l_s + l_5)^2 \dot{\theta}_2^2 + \dot{\theta}_2^2 \sum_{i=1}^{\infty} \phi_i(l_5) q_i(t)$$

$$\times \sum_{j=1}^{\infty} \phi_j(l_5) q_j(t) + \sum_{i=1}^{\infty} \phi_i(l_5) \dot{q}_i(t) \sum_{j=1}^{\infty} \phi_j(l_5) \dot{q}_j(t) \Bigg]$$

$$- m_L l_1 (l_s + l_5) \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_2 - \theta_1) + m_L$$

$$\times \Bigg[l_1 \sin(\theta_2 - \theta_1) \dot{\theta}_1 \dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(l_5) q_i(t)$$

$$+ (l_s + l_5) \dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(l_5) \dot{q}_i(t) \Bigg]$$

$$+ m_L \Bigg[-l_1 \cos(\theta_2 - \theta_1) \dot{\theta}_1 \sum_{i=1}^{\infty} \phi_i(l_5) \dot{q}_i(t) \Bigg] \tag{A8}$$

From (A2)–(A8), the total kinetic energy of the system is given by

$$T = \sum_{i=1}^{5} T_i + T_c + T_{\text{LED}}.$$

The eigenfunctions satisfy two orthogonality conditions (Hastings & Book 1986). The orthogonality conditions given by Hastings & Book (1986) are used to simplify the above expression for kinetic energy of the system. The orthogonality conditions are

$$\int_0^{l_5} \rho \phi_i \phi_j \, du = \begin{cases} 0, & \text{if } i \neq j, \\ \rho l_5, & \text{if } i = j, \end{cases}$$

$$\int_0^{l_5} EI \phi_i'' \phi_j'' \, du = \begin{cases} 0, & \text{if } i \neq j, \\ \omega_i^2 \rho l_5, & \text{if } i = j. \end{cases}$$

where $\phi_i(u)$ is the assumed-mode defined by (4). Let

$$d_1 = I_1 + m_1 l_{c1}^2 + I_3 + m_3 l_{c3}^2 + m_4 l_1^2 + m_5 l_1^2 + m_c l_1^2 + m_L l_1^2,$$

$$d_2 = I_2 + m_2 l_{c2}^2 + I_4 + m_3 l_2^2 + m_4 l_{c4}^2 + \tfrac{1}{3} m_5 l_5^2 + m_5 l_s^2 + m_5 l_s l_5$$
$$\qquad + m_c l_c^2 + m_L (l_s + l_5)^2,$$

$$d_3 = m_3 l_2 l_{c3} - \tfrac{1}{2} m_5 l_1 l_5 + m_4 l_1 l_{c4} - m_5 l_s l_1 - m_L l_1 (l_s + l_5) + m_c l_1 l_c.$$

It can be shown that

$$\int_0^{l_5} u \phi_i(u) \, du = 2/\beta_i^2,$$

and

$$\int_0^{l_5} \phi_i(u) \, du = \int_0^{l_5} \cosh \beta_i u - \cos \beta_i u - \sigma_i (\sinh \beta_i u - \sin \beta_i u) \, du$$

$$= 2\sigma_i/\beta_i.$$

$$\therefore \int_0^{l_5} (l_s + u) \phi_i(u) \, du = 2 \left(l_s \frac{\sigma_i}{\beta_i} + \frac{1}{\beta_i^2} \right) = \frac{\alpha_i}{\rho}, \quad \text{where } \alpha_i = 2\rho \left(l_s \frac{\sigma_i}{\beta_i} + \frac{1}{\beta_i^2} \right).$$

Using the orthogonality conditions and the relations mentioned above, the total kinetic energy of the system can be simplified as

$$T = \frac{1}{2} d_1 \dot{\theta}_1^2 + \frac{1}{2} d_2 \dot{\theta}_2^2 + d_3 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_2 - \theta_1) + \frac{1}{2} \rho l_5 \sum_{i=1}^{\infty} \dot{q}_i^2(t)$$

$$+ \frac{1}{2} \rho l_5 \dot{\theta}_2^2 \sum_{i=1}^{\infty} q_i^2(t) - 2\rho l_1 \dot{\theta}_1 \cos(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \left(\dot{q}_i(t) \frac{\sigma_i}{\beta_i} \right)$$

$$+ 2\rho l_1 \dot{\theta}_1 \dot{\theta}_2 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \left(q_i(t) \frac{\sigma_i}{\beta_i} \right) + \dot{\theta}_2 \sum_{i=1}^{\infty} (\dot{q}_i(t) \alpha_i) + \frac{1}{2} m_L$$

$$\times \left[\dot{\theta}_2^2 \sum_{i=1}^{\infty} \phi_i(l_5) q_i(t) \sum_{j=1}^{\infty} \phi_j(l_5) q_j(t) + \sum_{i=1}^{\infty} \phi_i(l_5) \dot{q}_i(t) \sum_{j=1}^{\infty} \phi_j(l_5) \dot{q}_j(t) \right]$$

$$
+ m_L \left[l_1 \sin(\theta_2 - \theta_1)\dot{\theta}_1\dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(l_5)q_i(t) + (l_s + l_5)\dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(l_5)\dot{q}_i(t) \right]
$$

$$
+ m_L \left[-l_1 \cos(\theta_2 - \theta_1)\dot{\theta}_1 \sum_{i=1}^{\infty} \phi_i(l_5)\dot{q}_i(t) \right]. \tag{A9}
$$

ce the manipulator moves in a horizontal plane, the total potential energy consists of
the elastic potential energy of link 5. The potential energy of the system is given by

$$
U = \frac{1}{2} \int_0^{l_5} EI \left(\frac{\partial^2 v}{\partial u^2} \right)^2 du,
$$

$$
U = \frac{1}{2} EI \int_0^{l_5} \left[\sum_{i=1}^{\infty} \phi_i''(u)q_i(t) \sum_{j=1}^{\infty} \phi_j''(u)q_j(t) \right] du,
$$

$$
U = \frac{1}{2} \rho l_5 \sum_{i=1}^{\infty} \omega_i^2 q_i^2(t). \tag{A10}
$$

ng the kinetic energy and potential energy expressions (A9) and (A10), the Lagrange
ation given by (A1) can now be used to find the equations of motion of the manipulator.

generalized coordinate, $z_1 = \theta_1$

$$
d_1\ddot{\theta}_1 + d_3 \cos(\theta_2 - \theta_1)\ddot{\theta}_2 - d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_2^2
$$

$$
- 2\rho l_1 \cos(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \ddot{q}_i(t)\frac{\sigma_i}{\beta_i} + 4\rho l_1\dot{\theta}_2 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \dot{q}_i(t)\frac{\sigma_i}{\beta_i}
$$

$$
+ 2\rho l_1 [\ddot{\theta}_2 \sin(\theta_2 - \theta_1) + \dot{\theta}_2^2 \cos(\theta_2 - \theta_1)] \sum_{i=1}^{\infty} q_i(t)\frac{\sigma_i}{\beta_i}
$$

$$
+ m_L l_1 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i(l_5)q_i\theta_2 + 2m_L l_1 \sin(\theta_2 - \theta_1)\dot{\theta}_2 \sum_{i=1}^{\infty} \phi_i(l_5)\dot{q}_i
$$

$$
+ m_L l_1 \cos(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i(l_5)q_i\dot{\theta}_2^2 - m_L l_1 \cos(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i(l_5)\ddot{q}_i = \tau_1. \tag{A11}
$$

generalized coordinate, $z_2 = \theta_2$

$$
d_2\ddot{\theta}_2 + d_3 \cos(\theta_2 - \theta_1)\ddot{\theta}_1 + d_3 \sin(\theta_2 - \theta_1)\dot{\theta}_1^2 + \sum_{i=1}^{\infty} \ddot{q}_i(t)\alpha_i
$$

$$
+ 2\rho l_1 [\ddot{\theta}_1 \sin(\theta_2 - \theta_1) - \dot{\theta}_1^2 \cos(\theta_2 - \theta_1)] \sum_{i=1}^{\infty} q_i(t)\frac{\sigma_i}{\beta_i} + \rho l_5\ddot{\theta}_2 \sum_{i=1}^{\infty} q_i^2(t)
$$

$$
+ 2\rho l_5\dot{\theta}_2 \sum_{i=1}^{\infty} q_i(t)\dot{q}_i(t) + m_L \sum_{i=1}^{\infty} \phi_i(l_5)q_i \sum_{j=1}^{\infty} \phi_j(l_5)q_j\ddot{\theta}_2
$$

$$
+ 2m_L \sum_{i=1}^{\infty} \phi_i(l_5)q_i \sum_{j=1}^{\infty} \phi_j(l_5)\dot{q}_j\dot{\theta}_2 + m_L l_1 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i(l_5)q_i\ddot{\theta}_1
$$

$$-m_L l_1 \cos(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \phi_i (l_5) q_i \dot{\theta}_1^2 + m_L (l_s + l_5) \sum_{i=1}^{\infty} \phi_i (l_5) \ddot{q}_i = \tau_2.$$

For generalized coordinate, $z_{k+2} = q_k$ $(k = 1, 2, 3, \ldots \infty)$

$$-\frac{2\sigma_i}{\beta_i} \rho l_1 \cos(\theta_2 - \theta_1) \ddot{\theta}_1 + \alpha_i \ddot{\theta}_2 + \rho l_5 \ddot{q}_i (t) + \rho l_5 (\omega_i^2 - \dot{\theta}_2^2) q_i (t)$$

$$-\frac{2\sigma_i}{\beta_i} \rho l_1 \dot{\theta}_1^2 \sin(\theta_2 - \theta_1) + m_L \phi_i (l_5)(\phi_1 (l_5) \ddot{q}_1 + \phi_2 (l_5) \ddot{q}_2)$$

$$+ m_L (l_s + l_5) \phi_i (l_5) \ddot{\theta}_2 - m_L l_1 \cos(\theta_2 - \theta_1) \phi_i (l_5) \ddot{\theta}_1$$

$$- m_L l_1 \sin(\theta_2 - \theta_1) \phi_i (l_5) \dot{\theta}_1^2 - m_L (\phi_1 (l_5) q_1 + \phi_2 (l_5) q_2) \phi_i l_5) \dot{\theta}_2^2 = 0.$$

The manipulator is actuated by 2 motors and therefore the motor dynamics must be modelled as well. The 2 motors used in this experiment are direct drive brush-less motors. Assuming only friction and viscous damping are present, the dynamic equations of the 2 motors can be written as

$$\begin{bmatrix} J_{m1} & 0 \\ 0 & J_{m2} \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} B_{ml} & 0 \\ 0 & B_{m2} \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} b_{m1} \text{sgn}(\dot{\theta}_1) \\ b_{m2} \text{sgn}(\dot{\theta}_2) \end{bmatrix} = \begin{bmatrix} \tau_{m1} \\ \tau_{m2} \end{bmatrix} - \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix},$$

$$(A12)$$

where J_{m_1}, J_{m_2} are the rotor inertia of the motors, B_{m_1}, B_{m_2} are the viscous damping coefficients, b_{m_1}, b_{m_2} are the friction coefficients, and τ_{m_1}, τ_{m_2} are the motor torques. From (A11) and (A12), the full dynamic model of the system can be obtained. This is given by

$$D(Q)\ddot{Q} + GQ + H(Q, \dot{Q}) + \mu(\dot{\theta}_1, \dot{\theta}_2) = PT_m,$$

$$Q = [\theta_1 \theta_2 q_1 q_2 \cdots q_\infty]^T, \quad T_m = [\tau_{m1} \tau_{m2}]^T,$$

$D(Q) =$

$$\begin{bmatrix} d_1 + J_{m1} & M_{12} & \left(\frac{-2\sigma_1}{\beta_1} \rho l_1 - m_L l_1 \phi_1 (l_5)\right) \times \cos(\theta_2 - \theta_1) & \left(\frac{-2\sigma_2}{\beta_2} \rho l_1 - m_L l_1 \phi_2 (l_5)\right) \times \cos(\theta_2 - \theta_1) & \cdots \\ M_{21} & \begin{matrix} d_2 + J_{m2} + \rho l_5 \sum_{i=1}^{\infty} q_i^2 \\ + m_L \sum_{i=1}^{\infty} \phi_i (l_5) q_i \\ \sum_{j=1}^{\infty} \phi_j (l_5) q_j \end{matrix} & \alpha_1 + m_L (l_s + l_5) \phi_1 (l_5) & \alpha_2 + m_L (l_s + l_5) \phi_2 (l_5) & \cdots \\ \left(\frac{-2\sigma_1}{\beta_1} \rho l_1 - m_L l_1 \phi_1 (l_5)\right) \times \cos(\theta_2 - \theta_1) & \alpha_1 + m_L (l_s + l_5) \phi_1 (l_5) & \rho l_5 + m_L \phi_1 (l_5)^2 & m_L \phi_1 (l_5) \phi_2 (l_5) & \cdots \\ \left(\frac{-2\sigma_2}{\beta_2} \rho l_1 - m_L l_1 \phi_2 (l_5)\right) \times \cos(\theta_2 - \theta_1) & \alpha_2 + m_L (l_s + l_5) \phi_2 (l_5) & m_L \phi_1 (l_5) \phi_2 (l_5) & \rho l_5 + m_L \phi_2 (l_5)^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & \rho l_5 \omega_1^2 & 0 & \cdots \\ 0 & 0 & 0 & \rho l_5 \omega_2^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\mu(\dot\theta_1, \dot\theta_2) = \begin{bmatrix} B_{m1}\dot\theta_1 + b_{m1}\,\mathrm{sgn}(\dot\theta_1) \\ B_{m2}\dot\theta_2 + b_{m2}\,\mathrm{sgn}(\dot\theta_2) \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \qquad P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \end{bmatrix},$$

$$H(Q, \dot Q) =$$

$$\begin{bmatrix} -d_3 \sin(\theta_2 - \theta_1)\dot\theta_2^2 + 4\rho l_1 \dot\theta_2 \sin(\theta_2 - \theta_1) \sum_{i=1}^{\infty} \dfrac{\sigma_i}{\beta_i}\dot q_i \\[1.2em] + 2m_L l_1 \sin(\theta_2 - \theta_1)\dot\theta_2 \sum_{i=1}^{\infty} \phi_i(l_5)\dot q_i \\[1.2em] + \left(2\rho l_1 \sum_{i=1}^{\infty} \dfrac{\sigma_i}{\beta_i} q_i + m_L l_1 \sum_{i=1}^{\infty} \phi_i(l_5) q_i \right) \cos(\theta_2 - \theta_1)\dot\theta_2^2 \\[1.2em] d_3 \sin(\theta_2 - \theta_1)\dot\theta_1^2 + 2\rho l_5 \dot\theta_2 \sum_{i=1}^{\infty} q_i\dot q_i + 2m_L \sum_{i=1}^{\infty} \phi_i(l_5) q_i \sum_{j=1}^{\infty} \phi_j(l_5)\dot q_j \dot\theta_2 \\[1.2em] - \left(2\rho l_1 \sum_{i=1}^{\infty} \dfrac{\sigma_i}{\beta_i} q_i + m_L l_1 \sum_{i=1}^{\infty} \phi_i(l_5) q_i \right) \cos(\theta_2 - \theta_1)\dot\theta_1^2 \\[1.2em] - \left(2\dfrac{\rho_1}{\beta_1}\rho l_1 + m_L l_1 \phi_1(l_5) \right) \sin(\theta_2 - \theta_1)\dot\theta_1^2 \\[1.2em] - \left(\rho l_5 q_1 + m_L \left(\sum_{i=1}^{\infty} \phi_i(l_5) q_i \right) \phi_1(l_5) \right) \dot\theta_2^2 \\[1.2em] - \left(2\dfrac{\rho_2}{\beta_2}\rho l_1 + m_L l_1 \phi_2(l_5) \right) \sin(\theta_2 - \theta_1)\dot\theta_1^2 \\[1.2em] - \left(\rho l_5 q_2 + m_L \left(\sum_{i=1}^{\infty} \phi_i(l_5) q_i \right) \phi_2(l_5) \right) \dot\theta_2^2 \\[1.2em] \vdots \end{bmatrix}$$

$$d_1 = I_1 + m_1 l_{c1}^2 + I_3 + m_3 l_{c3}^2 + m_4 l_1^2 + m_5 l_1^2 + m_c l_1^2 + m_L l_1^2,$$
$$d_2 = I_2 + m_2 l_{c2}^2 + I_4 + m_3 l_2^2 + m_4 l_{c4}^2 + \tfrac{1}{3} m_5 l_5^2 + m_5 l_s^2$$
$$\qquad + m_5 l_s l_5 + m_c l_c^2 + m_L (l_s + l_5)^2,$$

$$d_3 = m_3 l_2 l_{c3} - \tfrac{1}{2} m_5 l_1 l_5 + m_4 l_1 l_{c4} - m_5 l_s l_1 - m_L l_1 (l_s + l_5) + m_c l_1 l_c,$$

$$M_{12} = M_{21} = d_3 \cos(\theta_2 - \theta_1)$$
$$+ \left[2\rho l_1 \sum_{i=1}^{\infty} \frac{\sigma_i}{\beta_i} q_i + m_L l_1 \sum_{i=1}^{\infty} \phi_i(l_5) q_i \right] \sin(\theta_2 - \theta_1),$$

$$\alpha_i = 2\rho \left(l_s \frac{\sigma_i}{\beta_i} + \frac{1}{\beta_i^2} \right).$$

References

Aoustin Y et al 1994 Experimental results for the end-effector control of a single flexible robotic arm. *IEEE Trans. Control Syst. Technol.* 2: 371–381

Banavar R N, Dominic P 1995 An LQG/H-infinity controller for a flexible manipulator. *IEEE Trans. Control Syst. Technol.* 3: 409–416

Book W J 1984 Recursive Langrangian dynamics of flexible manipulor arms. *Int. J. Robotics Res.* 3: 87–101

Cannon R H, Schmitz E J 1984 Initial experiments on the end-point control of a flexible one-link robot. *Int. J. Robotics Res.* 3: 62–75

Craig R R Jr (ed.) 1981 *Structural dynamics: An introduction to computer methods* (New York: John Wiley & Sons)

De Luca A, Siciliano B 1993 Inversion based nonlinear control of robot arms with flexible links. *AIAA J. Guidance Control Dyn.* 16: 1169–1176

Gross E, Tomizuka M 1994 Experimental flexible beam tip tracking with a truncated series approximation to uncancellable inverse dynamics. *IEEE Trans. Control Syst. Technol.* 2: 382–391

Hastings G G, Books W J 1986 Verification of a linear dynamic model for flexible robotic manipulators. *Proc. of the IEEE Conference on Robotics and Automation* (Washington, DC: IEEE Comput. Soc. Press) pp 1024–1029

Khorrami F, Jain S, Tzes A 1994 Experiments on rigid body-based controllers with input pre-shaping for a two-link flexible manipulator. *IEEE Trans. Robotics Autom.* 10: 55–64

Lin L C, Yih T W 1996 Rigid model-based neural network control of flexible-link manipulators. *IEEE Trans. Robotics Autom.* 12: 595–601

Matsuno F, Fukushima S 1987 Feedback control of a flexible manipulator with a parallel drive mechanism. *Int. J. Robotics Res.* 6: 76–84

Mills J K, Lokhorst D M 1993 Control of robotic manipulators during general task execution — A discontinuous control approach. *Int. J. Robotics Res.* 12: 146–163

Moudgal V G, Passino K M, Yurkovich S 1994 Rule-based control for a flexible link robot. *IEEE Trans. Control Syst. Technol.* 2: 392–405

Sakawa Y, Matsuno F 1986 Modelling and control of a flexible manipulator with a parallel drive mechanism. *Int. J. Control* 44: 299–313

Siciliano B, Book W J 1988 A singular perturbation approach to control of lightweight flexible manipulators. *Int. J. Robotics Res.* 7: 79–90

Siciliano B, Prasad J V R, Calise A J 1992 Output feedback two-time-scale control of multilink flexible arms. *ASME J. Dyn. Syst. Meas. Control* 114: 70–77

Vandergrift M W, Lewis F L, Zhu S Q 1994 Flexible-link robot arm control by a feedback linearization/singular perturbation approach. *J. Robotic Syst.* 11: 591–603

D, Vidyasagar M 1991 Control of a class of manipulators with a single flexible link – part Feedback linearization. *ASME J. Dyn. Syst. Meas. Control* 113: 655–661

Q, Lewis F L, Hunt L 1994 Robust stabilization of the internal dynamics of flexible robots thout measuring the velocity of the deflection. *Proc. IEEE Conf. Decision and Control,* lando, pp 1811–1817

e actor–critic algorithm as multi-time-scale stochastic proximation

VIVEK S BORKAR* and VIJAYMOHAN R KONDA

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India

Abstract. The actor–critic algorithm of Barto and others for simulation-based optimization of Markov decision processes is cast as a two time scale stochastic approximation. Convergence analysis, approximation issues and an example are studied.

Introduction

rkov Decision Processes (MDPs) have been a popular paradigm for sequential decision cing under uncertainty. The traditional approach has been to write down the dynamic gramming equations appropriate for the problem at hand. Their solution yields the called value function for the problem. The optimal policy as a function of the state is a prescribed as the minimizer of a certain 'Hamiltonian' defined in terms of the value ction. This, in fact, gives a complete characterization of optimal (stationary) policies. ce everything hinges on computation of the value function, several iterative algorithms e been proposed for the same. These fall broadly into two classes : value iteration and icy iteration. An extensive account of these and related developments can be found in erman (1994).

Vhen applied to real problems, however, this scheme often runs into difficulties. The st notorious, of course, is the curse of dimensionality, caused by the typically very large e of the state space. An equally (if not more) difficult issue stems from the fact that the oretical analysis of MDPs presupposes exact knowledge of underlying stochastic dynics. Translated into real terms, this calls for accurate model selection and identification the relevant parameters. Though this can in principle be a separate, off-line statistical ercise, the computational overheads can be considerable.

This problem has been brought to the fore forcefully by some emerging applications artificial intelligence, e.g., in game playing machines and robotics (Barto *et al* 1995; erthi & Ravindran 1994). Here complexity of exact modeling and analysis is very high,

---

thor for correspondence

but, mercifully, that of simulation is often not so. To press this point further, consider a large interconnected system like a communication network. The dynamics evolves as per simple local update rules operating at individual constituent units and therefore is quite amenable to simulation on a parallel machine. But the overall dynamics can be extremely hard to analyse. This has prompted simulation-based algorithms which are akin to the traditional algorithms for computing the value function, but with one crucial difference: one replaces a computation using an exactly known transition probability function by an actual simulated transition as per the random mechanism that determines it. The algorithm is expected to 'see' the actual transition mechanism through an averaging affect. Mathematically speaking, these involve a marriage between the traditional algorithms for MDPs and stochastic approximation, a procedure in statistics with a long and illustrious history (see Benveniste *et al* 1990, for a comprehensive account of the latter).

These algorithms again fall into two broad classes. The first is the Q-learning algorithm of Watkins (1989) which has been extensively analysed (Watkins & Dayan 1992; Jaakola *et al* 1994; Tsitsiklis 1994). This can be viewed as a stochastic approximation version of the value iteration. The other strand consists of the actor–critic algorithm of Barto *et al* (1983) and its variants which may be viewed as stochastic approximation counterpart of policy iteration. Though some mathematical analysis is available in this case as well (Williams & Baird 1990), the situation is a little harder than Q-learning because of a certain inherent problem in policy iteration. In policy iteration, each update requires the computation of 'cost to go' function or 'value' function for a fixed policy. This in itself requires a 'value iteration' sandwiched between two policy updates, albeit a linear one (because the policy is fixed). One may view this as an algorithm with two loops, the inner one performing the linear value iteration for a fixed policy and the outer one updating the policy at the end of it. In practice, of course, the update of the outer loop cannot be kept waiting forever while the inner loop algorithm converges asymptotically to the desired cost to go function. Various *ad hoc* schemes have been proposed to ensure reasonable behaviour (Barto *et al* 1995). Our aim here is to propose a variant of the actor–critic algorithm based on some recent results on stochastic approximation with two time scales (Borkar 1996). The idea here is to operate the inner and outer loops with different step-size schedules, so that the inner loop moves on a faster effective time scale than the outer loop. This ensures that while the inner loop sees the current policy in the outer loop as quasi-static, the outer loop sees the value iteration of the inner loop as essentially equilibrated. This provides an actor–critic scheme that is asymptotically exact, at least in principle.

Having said all this, we should hasten to add that the simulation-based algorithms are not without problems. The first major problem is that many of these simulations call for a parallel, distributed implementation of the algorithm. This throws up issues like asynchronism and interprocessor communication delays. It is well-known that even simple, innocuous iterations that perform ideally in a centralised implementation can go haywire in a parallel, distributed environment (Chazan & Miranker 1969). Fortunately, a considerable body of work is now available on conditions under which one may still get the desired convergence (Bertsekas & Tsitsiklis 1989; Borkar 1994). We apply these ideas to the present algorithm to underscore conditions under which the desired convergence is preserved in a parallel distributed set-up.

The second problem is the familiar 'curse of dimensionality' which looms much larger simulation-based algorithms. This is because they involve iteration of vectors indexed state and action, not state alone, which increases the dimensionality many-fold (not mention that the algorithm perforce is restricted to finite state and action space case). us usually the algorithm must be accompanied by an approximation scheme to make it ctable. The traditional approach would be state aggregation whereby one clubs parts of te space into single meta-states. An alternative approach gaining currency is to approxi-te the value function directly (Schweitzer & Seidman 1985; Tsitsiklis & Van Roy 1996). is is appealing in view of the recently established function approximation properties of ural networks and good algorithms for neural network training that can be exploited here. vertheless, certain counterexamples in literature (Tsitsiklis & Van Roy 1996) suggest t the approach is not without its problems. We shall discuss these issues later in this ber.

The paper is organised as follows. The next section briefly reviews the MDP paradigm for counted cost problem and describes our variant of actor–critic algorithm associated with Section 3 recalls some key results concerning stochastic approximation, from Borkar)96) and Borkar (1994) respectively. These are used in § 4 for the convergence analysis of algorithm and its asynchronous version. Section 5 describes an approximation scheme sed on state aggregation. Section 6 presents a simulation example. Section 7 concludes th a brief discussion of further research issues.

This paper derives much from Borkar (1994; 1996) in terms of technique. Because of s, we have opted in favour of giving sketches of proofs instead of complete details. ving the latter would require reproducing the aforementioned references in totality, a nsiderable overhead in terms of length and mathematical abstraction. Needless to say, have pointed out the minor variations as and when needed. Complete mathematical tails can be found in Rao (1996).

MDPs and the actor–critic algorithm

e begin by recapitulating some well-known facts about MDPs. Puterman (1994) is a od general reference for the material.

Let $S = \{1, 2, \ldots, s\}$, $A = \{a_0, a_1, \ldots, a_r\}$ be prescribed finite sets and $p : S \times S \times$ $\rightarrow [0, 1]$ a map satisfying

$$p(i, j, a) \in [0, 1], \quad \sum_k p(i, k, a) = 1 \quad \forall i, j, a.$$

1 MDP (equivalently, a controlled Markov Chain) on state space S, with action space and transition probability function $p(\cdot, \cdot, \cdot)$, is an S-valued random process $X_n, n \geq 0$, tisfying

$$P(X_{n+1} = j / X_k, Z_k, k \leq n) = p(X_n, j, Z_n) \quad \forall n \geq 0,$$

here $\{Z_n\}$ is an A-valued 'control' process. If $\{Z_n\}$ is of the form $Z_n = v(X_n), n \geq 0$, r some map $v : S \rightarrow A$, we call $\{Z_n\}$, or, by abuse of terminology, the map v it-lf, a stationary policy. More generally, if for each n, Z_n is conditionally independent $X_m, Z_m, m < n$, given X_n, we call it a stationary randomised policy and identify

it with the map $\varphi : S \to \mathcal{P}(A)$ ($\mathcal{P}(\cdots) =$ the space of probability vectors on '$\cdots$') which gives the conditional law of Z_n given X_n. For $i \in S, a \in A$, let $\pi(i, a)$ denote the ath component of $\varphi(i)$. Under a stationary policy v (resp., a stationary randomised policy φ), $\{X_n\}$ is a time-homogeneous Markov chain with transition probabilities $[[p(i, j, v(i))]]$ (resp., $[[q(i, j, \varphi(i))]]$) where $q(i, j, \varphi(i)) = \sum_{a \in A} p(i, j, a)\pi(i, a)$. By a further abuse of terminology, we identify the stationary randomised policy φ with the vector $\pi = [\pi(i, a)]$, where the elements are ordered lexicographically. Note also that the class of stationary randomised policies contains the class of stationary policies, since the latter correspond to the case when each $\varphi(i)$ is concentrated at a single point in A.

We shall consider the infinite horizon discounted cost control problem. In this a discount factor $\alpha \in (0,1)$ and a running cost function $k : S \times A \to R$ are prescribed and the aim is to minimize over all admissible $\{Z_n\}$ the quantity

$$E\left[\sum_{n=0}^{\infty} \alpha^n k(X_n, Z_n)\right].$$

Define the 'value function' $V : S \to R$ by: For $i \in S$,

$$V(i) = \min E\left[\sum_{n=0}^{\infty} \alpha^n k(X_n, Z_n)/X_0 = i\right],$$

the minimum being over all admissible $\{Z_n\}$. It is known that V is the unique solution to the dynamic programming equations

$$V(i) = \min_a \left[k(i, a) + \alpha \sum_j p(i, j, a)V(j)\right], \quad i \in S.$$

Furthermore, any $v : S \to R$ satisfying: $v(i)$ attains the minimum in the rhs of the above, yields an optimal stationary policy, optimal for all initial conditions. In fact, $\{Z_n\}$ is optimal if and only if with probability one,

$$Z_n \in \arg\min(k(X_n, \cdot) + \alpha \sum_j p(X_n, j, \cdot)V(j)).$$

The key to the control problem therefore lies in finding $V(\cdot)$. Two standard approaches for this are value iteration and policy iteration. (A third approach to MDPs reduces them to linear programming problems. We do not consider this approach here.) The value iteration starts with an initial guess V_0 and iterates as per

$$V_{n+1}(i) = \min_a \left[k(i, a) + \alpha \sum_j p(i, j, a)V_n(j)\right], \quad i \in S,$$

for $n \geq 0$. Using Banach contraction mapping theorem, it is easy to show that $V_n \to V$ at an exponential rate.

The policy iteration scheme, on the other hand, starts with an initial guess $v : S \to A$ for an optimal policy and improves upon it iteratively as follows: At nth iteration
Step 1: Compute $V_n : S \to R$ defined by

$$V_n(i) = E\left[\sum_{m=0}^{\infty} \alpha^m k(X_m, Z_m)/X_0 = i\right], \quad i \in S,$$

expectation being under the stationary policy $Z_m = v_n(X_m)$, $m \geq 0$. This is done by
ving the linear equations

$$V_r(i) = k(i, v_n(i)) + \alpha \sum_j p(i, j, v_n(i)) V_n(j), \quad i \in S.$$

p 2: Find $v_{n+1} : S \to R$ by

$$v_{n+1}(i) \in \arg\min\left[k(i, \cdot) + \alpha \sum_j p(i, j, \cdot) V_n(j)\right], \quad i \in S.$$

One can show that the cost strictly decreases at each iterate as long as V_n is suboptimal,
suring convergence of $v_n(\cdot)$, $V_n(\cdot)$ to the optimal pair.
We now derive the appropriate 'simulation-based' version of this. There are some key
ferences between them which need to be underscored. The first, of course, is that we
lace each summation involving $p(i, \cdot, a)$ by a simulated transition as per that prob-
lity vector. In order for this to work, the algorithm should do some averaging. This
ensured by using an incremental version which makes only a small change in current
rates at each step, weighted by a stochastic approximation – like decreasing step-size.
condly, we operate with stationary randomised policies rather than stationary policies
that simple update equations for the probability vectors therein can be written. Finally,
linear system of step 1 is replaced by an iterative scheme for its solution before incor-
rating it into the simulation based scheme. This scheme is a 'stationary value iteration'
en by

$$V_n^{m+1}(i) = k(i, v_n(i)) + \alpha \sum_j p(i, j, v_n^m) V_n^m(i), \quad i \in S,$$

$m \geq 0$. This forms the 'inner loop' of the algorithm, wherein m is being updated for
ch fixed n. The equilibrium value $V_n(\cdot)$ to which $\{V_n^m(\cdot)\}$ will converge is then passed
to the outer loop for updating the policy. The crux of the algorithm we propose is to
nieve this two-tier structure by using two different time scales.
The 'centralized' version of our variant of the actor–critic algorithm is as follows. Let
$(n)\}$, $\{b(n)\}$ be decreasing sequences in $(0,1)$ satisfying

$$\sum_n a(n) = \sum_n b(n) = \infty \ , \ \sum_n a(n)^2, \sum_n b(n)^2 < \infty, \quad a(n) = o(b(n)).$$

$a_0 \in A$ and let P denote the projection of an r-vector onto the simplex $D =$
$1, \ldots, x_r] \mid x_i \geq 0, \ \forall i, \ \sum_i x_i \leq 1\}$. Let $\hat{\pi}(i)$ denote the r-vector $[\pi(i, a_1), \ldots, \pi(i, a_r)]$.
t e_a denote the r-vector whose components are indexed by elements of $A \setminus \{a_0\}$, with its
mponent indexed a as 1 and all other components equal to 0. The algorithm starts with
initial pair $V_0(\cdot) \in R^s$ and $\pi_0(i, a)$, $i \in S$, $a \in A \setminus \{a_0\}$, and iterates according to

$$V_{n+1}(i) = (1 - b(n)) V_n(i) + b(n)[\bar{k}(i, \varphi_n(i)) + \alpha V_n(\xi_n(i))],$$

$i \in S$, and for each $i \in S$, $a \in A \setminus \{a_0\}$,

$$\hat{\pi}_{n+1}(i) = P\left(\hat{\pi}_n(i) + a(n)\left(\sum_{a' \in A \setminus \{a_0\}} (V_n(i) - k(i, a')\right.\right.$$

$$\left.\left. - \alpha V_n(\eta_n(i, a')))\pi_n(i, a')e_{a'} + \phi'(n)\right)\right)$$

$$\pi_{n+1}(i, a_0) = 1 - \sum_{a \neq a_0} \pi_{n+1}(i, a),$$

for $n \geq 0$, with $\varphi(i) \triangleq \pi_n(i, \cdot) \in \mathcal{P}(A)$ and $\bar{k}(i, \varphi) = \sum_a \pi(i, a)k(i, a)$ for $\varphi \in \mathcal{P}(A)$. Furthermore, $\xi_n = [\xi_n(1), \ldots, \xi_n(s)]$, $\eta_n = [[\eta_n(i, a)]]$ are resp. S^s, $S^{s \times r}$-valued random variables conditionally independent of each other given ξ_i, η_i, $i < n$, with the corresponding conditional distributions equal to

$$\prod_{j=1}^{s} q(j, ., \varphi_n(j)) \quad \text{and} \quad \prod_{j \in S} \prod_{a \in A} p(j, ., a)$$

respectively. The reader may verify that this is in confirmation with our verbal description earlier. A small modification, however, is warranted. If any $\pi_n(i, \cdot)$ is on a face of the simplex $\mathcal{P}(A)$, it will remain there thereafter. To avert this, a small diminishing noise $\phi'(n)$, with a Lebesgue continuous law, is added to push it away from the stable manifold of the unstable equilibrium points on the face of the simplex.

The distributed, asynchronous version is more complicated and needs additional notation and assumptions. To start with, let I_1, I_2 be sets of subsets of S, $S \times A \setminus \{a_0\}$ resp. that together cover resp. S, $S \times A \setminus \{a_0\}$. Let $\{Y_n\}$, $\{Z_n\}$ be resp. I_1-, I_2- valued processes with the interpretation: Y_n is the set of $i \in S$ such that $V_n(i)$ gets updated at time n and Z_n is the set of $(i, a) \in S \times A \setminus \{a_0\}$ such that $\pi_n(i, a)$ gets updated at time n. We impose on these processes the condition: There exists a deterministic $\Delta > 0$ such that with probability one,

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} I\{i \in Y_m\} \geq \Delta, \ i \in S,$$

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} I\{(i, a) \in Z_m\} \geq \Delta, \ i \in S, a \in A.$$

This ensures that the distributed asynchronous version updates all components comparably often in a precise sense. (See Borkar 1994 for a graph theoretic sufficient condition for the above to hold.)

Secondly, we introduce delays $\tau_n(i, j)$, $\bar{\tau}_n(i, a, j)$, $\hat{\tau}(i, a)$ taking values in $[0, 1, \ldots, \min(n, N)]$, $N \geq 0$, with the assumption: $\tau_n(i, i) = 0 \ \forall i, n$. The idea is: Each component of the iteration is updated by a fixed processor, which receives the updates from other processors with random but bounded interprocessor communication delays. Thus the processor updating $V_n(i)$ receives at time n, $V_{n-\tau_n(i, j)}(j)$ and not $V_n(j)$. Similarly, the processor updating $\pi_n(i, a)$ receives at time n, $V_{n-\bar{\tau}_n(i, a, j)}(j)$ and not $V_n(j)$. Finally, the

essor updating $V_n(i)$ at time n receives $\pi_{n-\hat{\tau}_n(i,a)}(i, a)$ and not $\pi_n(i, a)$. Assume that
$i, j), \bar{\tau}_n(i, a, j), \hat{\tau}_n(i, a)\}$ are independent of $\xi_m,\ \eta_m\ m < n$ for each n. The latter
defined as before, except that the conditional law of ξ_n given $\xi_m,\ \eta_m,\ m < n$ gets
aced by $\prod_{j \in S} q(j, \cdot, \varphi_n(j))$ where

$$\varphi_n(i) = \pi_{n-\hat{\tau}(i, \cdot)}(i, \cdot),\ i \in S.$$

hould be remarked that the boundedness condition on delays can be replaced by a mild
ditional moment bound as in Borkar 1994 at the expense of additional technicalities.)
inally, introduce for $n \geq 0$,

$$\nu_1(i, n) = \sum_{m=0}^{n} I\{i \in Y_m\},\ i \in S,$$

$$\nu_2(i, a, n) = \sum_{m=0}^{n} I\{(i, a) \in Z_m\},\ i \in S,\ a \in A.$$

e that the processor updating $V_n(i)$ (resp. $\pi_n(i, a)$) knows $\nu_1(i, n)$ (resp. $\nu_2(i, a, n)$) at
n (that being the number of updates he has performed till then), even when he does
know n, i.e., the 'global clock'.
he distributed, asynchronous version of the algorithm then is: for $i \in S,\ a \in A \setminus \{a_0\}$,

$$V_{n+1}(i) = V_n(i) + b(\nu_1(i, n))[\bar{k}(i, \varphi_n(i))$$
$$+ \alpha V_{n-\tau_n(i, \xi_n(i))}(\xi_n(i)) - V_n(i)]I\{i \in Y_n\},$$

$$\hat{\pi}_{n+1}(i) = P\left(\hat{\pi}_n(i) + \sum_{a' \in A \setminus \{a_0\}} a(\nu_2(i, a, n))((V_n(i) - k(i, a)\right.$$
$$- \alpha V_{n-\bar{\tau}(i, a, \eta_n(i, a))}(\eta_n(i, a)))\pi_n(i, a)$$

$$\left. + \phi(n))I\{(i, a) \in Z_n\}e_{a'}\right),$$

$$\pi_{n+1}(i, a_0) = 1 - \sum_{a \neq a_0} \pi_{n+1}(i, a),$$

0 and $\{\phi(n)\}$ is a random sequence converging to zero. The role of this sequence is
e as that of $\phi(n)$ in synchronous algorithm. For this algorithm, we shall impose the
owing additional restrictions on $\{a(n)\},\ \{b(n)\}$: Let $\{c(n)\}$ denote $\{a(n)\}$ or $\{b(n)\}$.
n

There exists $r \in (0, 1)$ such that $\sum_n c(n)^{1+r} < \infty$.

For $x \in (0, 1)$, $\sup_n c([xn])/c(n) < \infty$, where $[\cdots]$ stands for 'the integer part of
$\cdots$'.

For $x \in (0, 1)$ and $A(n) = \sum_{m=0}^{n} c(m)$, $A([yn])/A(n) \to 1$ uniformly in $y \in [x, 1]$.

xamples of $\{c(n)\}$ satisfying the above are: $1/n,\ 1/n \ln(n),\ \ln(n)/n$ etc., with modifi-
on for $n = 0, 1$ where needed.
Ve shall analyse these algorithms in § 4 after a brief review of some relevant topics in
chastic approximation in the next section.

This section briefly recalls some recent results in stochastic approximation algorithms needed for our work. The stochastic approximation algorithm in its simplest form is the d-dimensional iteration

$$X(n+1) = X(n) + a(n)(h(X(n)) + M(n)), \ n \geq 0, \tag{1}$$

where $\{a(n)\}$ is as before and $\{M(n)\}$ is a sequence of integrable random variables satisfying

$$E[M(n)/X(m), \ m \leq n, \ M(m), \ m < n] = 0, \ n \geq 0. \tag{2}$$

The convergence of this algorithm to a desired limit is usually established by first establishing separately that with probability one,

$$\sup_n | X(n) | < \infty, \ \sum_n a(n)M(n) < \infty. \tag{3}$$

Given these, one way to analyse its asymptotic behaviour is by showing that it asymptotically tracks the ordinary differential equation (ODE) given by

$$\dot{x}(t) = h(x(t)).$$

Assume h to be Lipschitz with linear growth, ensuring that this ODE has a unique solution for any initial condition, defined for all $t \geq 0$. Suppose this ODE has a globally asymptotically stable attractor J. Then by converse Liapunov theorem (see, e.g., Yoshizawa 1966) there exists a continuously differentiable $V : R^d \rightarrow R^+$ satisfying $V(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$ and $\nabla V(x) \cdot h(x) < 0$ for $x \notin J$. Now, given $T, \delta > 0$, call a bounded measurable function $y(\cdot) : R^+ \rightarrow R^d$ a (T, δ)-perturbation of this ODE if there exist $0 = T_0 < T_1 < T_2 < \cdots T_i \uparrow \infty$, such that $T_{j+1} \geq T_j + T$ and there exist solutions $x^j(\cdot)$ of the ODE on each interval $[T_j, T_{j+1}]$ such that

$$\sup_{t \in [T_j, T_{j+1}]} \|x^j(t) - y(t)\| < \delta.$$

Then one has:

Lemma 1 (Hirsch 1989). *For every $\epsilon > 0$, $T > 0$, there exists a $\delta_0 > 0$ such that for any $\delta \in (0, \delta_0)$, every (T,δ)-perturbation of the ODE converges to the ϵ-neighbourhood of J.*

The idea of the proof is: The Liapunov function V must decrease by a minimum positive quantity along every $x^j(\cdot)$ that does not intersect the ϵ-neighbourhood of J and therefore along the corresponding patch of $y(\cdot)$. This can happen for at most finitely many consecutive j's, so $y(\cdot)$ must eventually intersect this neighbourhood. Analogous considerations show that it cannot move away too much once it has done so. (See the appendix of Borkar 1996 for details.)

The convergence analysis of the stochastic approximation algorithm now hinges on the following 'time-scaling' argument: Let $t(0) = 0$, $t(n) = \sum_{m=0}^{n-1} a(m)$, and pick $m(n)$ according to: $m(0) = 0$, $m(n) = \min\{t(k) \mid t(k) \geq t(m(n)) + T\}$. Set $T_j = t(m(j))$,

0. Define $y(t)$, $t \geq 0$, by: $y(t(m)) = X(m)$, $m \geq 0$, with linear interpolation $t(m), t(m+1)]$, $m \geq 0$. Let $x^j(\cdot)$ be the solution of the ODE on $[T_j, T_{j+1}]$ with $_j) = y(T_j)$, $j \geq 0$. Then $y(\cdot)$ on $[T_j, T_{j+1}]$ may be viewed as an Euler approximation e ODE with a nonuniform but decreasing (with j) step-size, modulo an error term due $M(n)\}$ that also becomes asymptotically negligible thanks to (3). The above lemma applies 'eventually' (i.e., for sufficiently large j) for each $\epsilon \geq 0$, ensuring $X(n) \to J$ probability one.

ow consider a 'two time-scale' variant of the basic algorithm:

$$X(n+1) = X(n) + a(n)(F(X(n), Y(n)) + M(n+1)),$$

$$Y(n+1) = Y(n) + b(n)(G(X(n), Y(n)) + M'(n+1)),$$

re F, G are Lipschitz with linear growth and $M(n)$, $M'(n)$ are integrable random bles uncorrelated with the past (i.e., satisfying (2)) and $a(n) = o(b(n))$. Suppose probability one, the following hold:

$$\sup_n | X(n) | < \infty, \quad \sum_n a(n)M(n) < \infty,$$

$$\sup_n | Y(n) | < \infty, \quad \sum_n a(n)M'(n) < \infty.$$

lso suppose that for each x, the ODE,

$$\dot{y}(t) = G(x, y(t)), \tag{4}$$

a unique globally asymptotically stable equilibrium point $\lambda(x)$ where $\lambda(\cdot)$ is Lipschitz inuous and the ODE,

$$\dot{x}(t) = F(x(t), \lambda(x(t))), \tag{5}$$

a unique globally asymptotically stable attractor J.

orem 1. *With probability one, $(X(n), Y(n)) \to \{(x, \lambda(x)) \mid x \in J\}$.*

proof can be found in Borkar (1996) for the case when J is a singleton and extends ly to the more general case. The idea is to mimic the above time scaling argument first $b(n)$ (i.e., $t(n) = \sum_{i=0}^n b(i)$), so that the interpolated trajectories track the ODE

$$\dot{x}(t) = 0, \quad \dot{y}(t) = G(x(t), y(t)),$$

then again with $a(n)$ (i.e., $t(n) = \sum_{i=0}^n a(i)$), so that the interpolated $\{X(n)\}$ tracks Lemma 1 is used in each case in the obvious manner. Thus the fast component $\{Y(n)\}$ the slow component $\{X(n)\}$ as quasi-static, while the slow component sees the fast ponent as 'essentially equilibrated'.

he distributed, asynchronous version of the algorithm (1) is as follows: If $X_i(n)$ is the component of the vector $X(n)$, it is updated as per

$$X_i(n+1) = X_i(n) + a(\nu(i, n))(h(X_1(n - \tau_n(i, 1)), \ldots, X_d(n - \tau_n(i, d)))$$

$$+ M(n))I\{i \in Y_n\}$$

For this case, the rather intricate analysis of Borkar (1994) shows that a suitably inter-polated version of $\{X(n)\}$, if bounded, tracks the ODE

$$\dot{x}(t) = \frac{1}{d} h(x(t)).$$

The scalar $1/d$ up front amounts to linear time scaling that does not alter the qualitative behaviour. Thus $X(n) \to J$ with probability one as before. (It should be remarked that the algorithm considered in Borkar 1994 is slightly more restrictive than (1) or (5), but the same arguments go through nevertheless.)

4. Convergence analysis

In this section we shall adapt the ideas of the preceding section for the convergence analysis of the actor–critic algorithm proposed in § 2. As already mentioned earlier, only a sketch of the proofs will be given.

Define the map $F : \mathcal{P}(A)^s \times R^s \to R^s$ by $F(\cdot, \cdot) = [F_1(\cdot, \cdot), \ldots, F_s(\cdot, \cdot)]^T$ where $F_i(z, x) = \sum_a z(i, a) k(i, a) + \alpha \sum_j \sum_a z(i, a) p(i, j, a) x_j$ for $x = [x_1, \ldots, x_s]^T$ and $z = [[z(i, a)]]$, $i \in S$, $a \in A$ with $z(i, \cdot) \in \mathcal{P}(A)$ $\forall i$. Also define the following norms on R^s

$$\|x\|_p = \left(\frac{1}{s} \sum_{i=1}^{s} |x|^p \right)^{\frac{1}{p}}, \quad p \in (1, \infty),$$

$$\|x\|_\infty = \max_i |x_i|.$$

Then it is easily verified that for each fixed z, the following contraction condition holds:

$$\|F(z, x) - F(z, y)\|_\infty \le \alpha \|x - y\|_\infty, \quad x, y \in R^s.$$

The first iteration of the centralised algorithm can now be written as

$$V_{n+1} = V_n + b(n)(F(\pi_n, V_n) - V_n) + b(n) M(n), \quad n \ge 0 \tag{6}$$

for suitably defined $M(n)$ which will satisfy (2) (with $X(m) \stackrel{\triangle}{=} V_m$).

Lemma 2. With probability one, the iterates of the centralized or asynchronous algorithm remain bounded.

Proof. We consider the asynchronous case, as the centralized case is a special case thereof. The iterates of $\{\pi_n\}$ are bounded anyway because of the projection P onto a bounded set. That V_n remain bounded with probability one follows exactly as in theorem 1, pp. 190–191, of Tsitsiklis (1994). (In Tsitsiklis 1994, the algorithm slightly differs from (6) insofar as $F(\pi_n, V_n)$ would be replaced by $\bar{F}(V_n)$ for some $\bar{F}$ satisfying (say) $\|\bar{F}(x) - \bar{F}(y)\|_\infty \le \alpha \|x - y\|_\infty$. Nevertheless, exactly the same proof applies.) □

In turn, this implies that $\{M(n)\}$, which is defined in terms of V_n, remains bounded with probability one. A direct calculation shows that in addition

$$\psi(n) \stackrel{\triangle}{=} E[M(n)^2 / V_m, \pi_m, \tau_m(i, j), \bar{\tau}_m(i, a, j), \hat{\tau}_m(i, a), Y_m, Z_m,$$
$$m \le n, i, j \in S, a \in A]$$

main bounded with probability one. Since $\sum_n b(n)^2 < \infty$, we then have $\sum_n b(n)^2 \psi(n)^2$
∞ with probability one. The property (2) renders the process

$$\sum_{i=0}^{n} b(i) M(i), n \geq 0,$$

martingale and the foregoing then ensures that its quadratic variation process (see Neveu
75, for a definition) is convergent with probability one. Proposition VII-3-(c), pp. 149–
0 of Neveu (1975) then ensures that

$$\sum_{i=0}^{\infty} b(i) M(i)$$

convergent, and hence finite with probability one. This, in conjunction with lemma 2
ove, completes our verification of (3) for iteration (6).

These considerations will enable us to use the results of the preceding section. Prior
doing so, we look at the expected ODE limits. The first one to be considered is the
unterpart of (4) in our set-up, which is

$$\dot{x}(t) = F(\pi, x(t)) - x(t) \tag{7}$$

F as above and a fixed $\pi \in \mathcal{P}(A)^s$.

mma 3. *Equation* (7) *has a unique globally asymptotically stable equilibrium point* $\bar{V}_\pi$
ven by

$$\bar{V}_\pi(i) = E\left[\sum_{m=0}^{\infty} \alpha^m k(X_m, Z_m)/X_0 = i \right],$$

e expectation being with respect to the stationary randomised policy π.

oof. Standard contraction mapping arguments show that $F(\pi, \cdot)$ has a unique fixed point
d usual dynamic programming type arguments show that $\bar{V}_\pi$ defined above must be it.
ow for $p \in (1, \infty)$, explicit differentiation leads to

$$\frac{d}{dt}\|x(t) - \bar{V}_\pi\|_p = \|x(t) - \bar{V}_\pi\|_p^{1-p}(-\|x(t) - \bar{V}_\pi\|_p$$

$$+ \frac{1}{s} \sum_{i=1}^{s} (x_i(t) - \bar{V}_\pi(i))^{p-1}(F_i(\pi, x(t)) - F_i(\pi, \bar{V}_\pi))$$

$$\leq -\|x(t) - \bar{V}_\pi\|_p + \|F(\pi, x(t)) - F(\pi, \bar{V}_\pi)\|_p$$

here the inequality follows from an application of Hölder's inequality. Let $p \to \infty$.
sing continuity of $p \to \|x\|_p$ on $[1, \infty]$ and the contraction property of $F(\pi, \cdot)$ under
e $\|\cdot\|_\infty$-norm, we have

$$\frac{d}{dt}\|x(t) - \bar{V}_\pi\|_\infty \leq -(1-\alpha)\|x(t) - \bar{V}_\pi\|_\infty,$$

Note that since V_π is characterized by the linear system of equations $F(\pi, V_\pi) = V_\pi$, it depends smoothly on the coefficients thereof, hence on π. In particular, the map is Lipschitz on $\mathcal{P}(A)^s$.

Define a vector field $G(\cdot)$ on $\mathcal{P}(A)^s$ by: $G(\cdot) = [[G_{ia}(\cdot)]]$ for $i \in S$, $a \in A$, where

$$G_{ia}(\pi) = \pi(i, a) \left(V_\pi(i) - k(i, a) - \alpha \sum_j p(i, j, a) V_\pi(j) \right).$$

The second ODE (the counterpart of (5)) that we consider is

$$\dot{\pi}(t) = G(\pi(t)). \tag{8}$$

It is easy to verify that this ODE will always remain in $\mathcal{P}(A)^s$ if $\pi(0)$ is. Furthermore, it will converge to equilibrium points, the latter being the set of those $\pi \in \mathcal{P}(A)^s$ that satisfy: for each (i, a), either $\pi(i, a) = 0$ or $k(i, a) + \alpha \sum_j p(i, j, a) V_\pi(j) = V_\pi(i)$. A little thought shows that the stable equilibrium points are those for which, for each (i, a), either $\pi(i, a) = 0$ or

$$V_\pi(i) = \min_a \left(k(i, a) + \alpha \sum_j p(i, j, a) V_\pi(j) \right).$$

Since not all $\pi(i, a)$ can be zero for a given i, the above equality holds for all i. Hence the only stable equilibrium points are the π for which V_π equals V. These are precisely the optimal stationary randomised policies.

Consider the centralized algorithm.

Theorem 2. *With probability one, the centralized algorithm converges to $\{(V, \pi) \mid \pi$ is an optimal stationary randomised policy$\}$.*

Proof. We can adopt the 'two-time' scale results of the preceding section with one difference. The difference here is the presence of the projection operator P in the algorithm. This, however, is a standard feature of many stochastic approximation algorithms and the theory thereof is well-understood; see, e.g., Kushner & Clark (1978). As in the theorem 5.3.1, p. 191, of Kushner & Clark, we can conclude that the limiting ODE tracked by a suitably interpolated version of $\{\pi_n\}$ (and hence by $\{\pi_n\}$ itself) is given by the $R^{s \times (r-1)}$-dimensional ODE

$$\dot{\pi}(t) = \tilde{P}(\bar{G}(\pi(t))),$$

where $\bar{G}_{ia}(\pi) = G_{ia}([\pi[1, \cdot]; 1 - \sum_{b \neq a_0} \pi(1, b); \pi(2, a); 1 - \sum_{b \neq a_0} \pi(2, b); \cdots ; \pi(s, a); 1 - \sum_{b \neq a_0} \pi(s, b)])$, $j \in S$, and $\tilde{P}(h(\cdot))$, for a vector field $h(\cdot)$ on $\mathcal{P}(A)^s$ is defined by

$$\tilde{P}(h(y)) = \lim_{0 < \delta \to 0} \frac{P(y + h(y)\delta) - P(y)}{\delta}.$$

If the limit is non-unique, one considers the set of all limit points and treats the above as a differential inclusion rather than an ODE. Fortunately for us, $\tilde{P}(\bar{G}(\pi))$ is well-defined and in fact equals $G(\pi)$, as can be easily verified. Thus the limiting ODE is

$$\dot{\pi} = G(\pi(t)).$$

now augment $\pi(t)$ by adjoining s additional components $\pi(i, a_0) = 1 - \sum_a \pi(i, a)$, $i \in$
Denote the enlarged vector by $\pi(t)$ again by abuse of notation. It is easy to verify that
s redefined, $\pi(\cdot)$ satisfies (8). The convergence for the two time scale stochastic algo-
ims can now be invoked to claim that with probability one, $\{\pi_n\}$ converges to the set of
ilibrium points of (8). The final step is to show that it will in fact converge to the set of
ple equilibrium points of (8) with probability one (Pemantle (1990), see also Brandière
Duflo (1996)). To show this, all we need is that the set of points that get attracted to an
stable equilibrium point have zero probability. It is easy to see that Lebesgue measure of
stable manifold of an unstable equilibrium point is zero. Thus it is enough if the noise
1) has positive density with respect to Lebesgue measure in a shrinking neighbourhood
the origin. The claim follows. □

eorem 3. *The claims of theorem 2 also hold for the decentralized algorithm under
ditional hypotheses stipulated in § 3.*

This follows simply by combining the foregoing with the scheme of Borkar (1994). This
mpletes our convergence analysis of the algorithm.

Approximation issues

mentioned in the introduction, in many applications the 'curse of dimensionality' forces
e to interface the above and other simulation-based algorithms with an approximation
heme. Here we sketch a scheme based on state aggregation, adapted from Tsitsiklis
Van Roy (1994) where it is proposed as a 'look-up table scheme' in the context of
learning.

The idea is to partition the state space S into disjoint nonempty subsets $S_1, S_2, \ldots, S_m$,
ch identified as an aggregated state. For each j, let $\beta_j(\cdot)$ be a probability vector on S_j
th $\beta_j(i) > 0$, $\forall i \in S_j$. The centralized version of the algorithm runs as follows: At
ch n, generate for each j, $1 \leq j \leq m$, random variables $(X_j^n, \xi_j^n, \eta_{ja}^n, \ a \in A)$ with law

$$P(X_j^n = i_j, \xi_j^n = k_j, \eta_{ja}^n = x_{aj}, \ a \in A, j = 1, 2, \ldots, m)$$

$$= \prod_{j=1}^{m} \beta_j(i_j) q(i_j, k_j, \varphi_n(j)) \prod_{a \in A} p(i_j, x_{aj}, a)$$

dependently of the past, where $\varphi_n(j) = \pi_n(j, .)$. Define $\bar{\eta}_{ja}^n \in \{1, 2, \ldots, m\}$ by $\bar{\eta}_{ja} = k$
$\eta_{ja}^n \in S_k$. The iteration at time n is: for each j, $1 \leq j \leq m$,

$$V_{n+1}(j) = (1 - b(n))V_n(j) + b(n)(\bar{k}(X_j^n, \varphi_n(j)) + \alpha V_n(\xi_j^n)),$$

$$\hat{\pi}_{n+1}(j) = P\bigg(\hat{\pi}_n(j) + a(n)\bigg(\sum_{a \in A \setminus \{a_0\}} V_n(j) - \bar{k}(X_j^n, a)$$

$$-\alpha V_n(\bar{\eta}_{ja}))\pi_n(j, a)e_0 + \phi(n)\bigg)\bigg),$$

$$\pi_{n+1}(j, a_0) = 1 - \sum_{a \neq a_0} \pi_{n+1}(j, a),$$

where P is the projection onto the appropriate r-dimensional simplex.

An analysis similar to the foregoing shows that with probability one, these iterates converge to $(V^*(\cdot), \pi^*(\cdot, \cdot))$, where for $j \in \{1, 2, \ldots, m\}$,

$$V^*(j) = \min_a \left[\sum_{i \in S_j} \beta_j(i)[k(i, a) + \alpha \sum_{l=1}^{m} V(l) \sum_{x \in S_l} p(i, l, a)] \right]$$

and support $(\pi^*(j, .)) \subset \arg\min$ of the rhs above. For error estimates for such schemes, see Tsitsiklis & Van Roy (1994)

A distributed version of this approximate scheme along the lines of the foregoing can also be analysed accordingly.

6. Simulation experiments

To demonstrate the convergence of the proposed algorithm we took two examples. The MDPs of these two examples are similar to those used to model admission control into an $M/M/1$ queue and routing to parallel $M/M/1$ queues respectively (Walrand 1988). In both these examples, in any iteration, only one component of V and one component of π are updated. The probability of update of a component in an iteration is equal for all components and is independent of everything else. Since $\pi_n(i, a)$ cannot change once it becomes zero, a small noise is added whenever it is sufficiently close to zero. The deterministic value iteration was used to find out optimal value function and policy. The algorithm was started with value function being identical to zero, and with all the randomised policy probabilities equal. The examples are described below. The notation used here is the one introduced in § 2. Interprocessor communication is assumed to be instantaneous (i.e., no delays).

6.1 *Example 1*

$$S = \{0, 1, \ldots, N\},$$
$$A = \{0, 1\},$$
$$p(i, j, a) = \begin{cases} 0, & \text{if } j < \max(j - 1, 0) \text{ or } j > i + 1, \\ \lambda, & \text{if } j = \min(i + 1, N) \text{ and } a = 0, \\ & \text{or } j = i \text{ and } a = 1, \\ 1 - \lambda, & \text{if } j = \max(i - 1, 0), \end{cases}$$
$$k(i, a) = \begin{cases} i, & \text{if } a = 0, \\ i + c, & \text{if } a = 1. \end{cases}$$

For the purpose of simulation we took:

$$N = 10, \quad a(n) = (\{\lfloor \tfrac{n}{100} \rfloor + 1\}/n)^{\frac{2}{3}}$$
$$\lambda = 0.65,$$
$$c = 10, \quad b(n) = \{\lfloor \tfrac{n}{100} \rfloor + 1\}/n,$$
$$\alpha = 0.99.$$

The optimal value function and policy were:

$$V = \begin{bmatrix} 669.0261 & 679.4227 & 694.0252 & 711.0231 & 730.3492 & 751.9383 \\ & 775.7268 & 801.6532 & 820.9869 & 831.7235 & 836.4438 \end{bmatrix},$$

$$\pi(\cdot, 0) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$
$$\pi(i, 1) = 1 - \pi(i, 0) \quad i \in S.$$

After 2×10^6 iterations the result of our algorithm was

$$\hat{V} = \begin{bmatrix} 663.2563 & 673.3967 & 687.6990 & 708.0571 & 725.9087 & 748.0945 \\ & 775.7268 & 796.7382 & 820.2483 & 830.5291 & 835.0053 \end{bmatrix},$$

$$\hat{\pi}(\cdot, 0) = \begin{bmatrix} 1 & 0.9986 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9871 & 1 & 0.9955 \end{bmatrix},$$
$$\hat{\pi}(i, 1) = 1 - \hat{\pi}(i, 0) i \in S.$$

6.2 *Example 2*

$$S = \{(i, j), \ 0 \le i \le N, \ 0 \le j \le N\},$$
$$A = \{0, 1\},$$

$$p((i, j), (k, l), a) = \begin{cases} \lambda, & \text{if } k = \min(i + 1, N) \text{ and } j = l \\ & \quad \text{and } a = 0 \\ & \quad \text{or } k = i \text{ and } l = \min(j + 1, N) \\ & \quad \text{and } a = 1, \\ \mu_1, & \text{if } k = \max(i - 1, 0) \text{ and } l = j, \\ 1 - \lambda - \mu_1, & \text{if } k = i \text{ and } l = \max(j - 1, 0), \\ 0, & \text{otherwise,} \end{cases}$$

$$k((i, j), a) = c_1 i + c_2 j.$$

For the purpose of simulation we took:

$$\begin{aligned} N &= 10, & a(n) &= (\{\lfloor \tfrac{n}{100} \rfloor + 1\}/n)^{\frac{2}{3}} \\ \lambda &= 0.65, & b(n) &= 0.1(\{\lfloor \tfrac{n}{100} \rfloor + 1\}/n) \\ c_1 &= 10, & c_2 &= 15, \\ \alpha &= 0.99, & \lambda &= 0.5, \\ \mu_1 &= 0.3, & \mu_2 &= 0.2. \end{aligned}$$

The optimal value function and policy were:

$$[[V(i, j)]] = 10^3 \times \begin{bmatrix} 3.3602 & 3.4448 & 3.5978 & 3.8137 & 4.0899 & 4.2539 \\ 3.4281 & 3.5179 & 3.6710 & 3.8863 & 4.1618 & 4.3018 \\ 3.5277 & 3.6183 & 3.7696 & 3.9833 & 4.2578 & 4.3875 \\ 3.6380 & 3.7171 & 3.8643 & 4.0761 & 4.3496 & 4.4937 \\ 3.7172 & 3.7922 & 3.9370 & 4.1474 & 4.4201 & 4.6036 \\ 3.7590 & 3.8328 & 3.9766 & 4.1864 & 4.4588 & 4.7032 \end{bmatrix},$$

$$[[\pi((i,j),0)]] = \begin{bmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0 \\ 0 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0 \\ 0 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0 \\ 0\text{ to }1 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 0\text{ to }1 \end{bmatrix}.$$

After 10^7 iterations the result of our algorithm was

$$[[\hat{V}(i,j)]] = 10^3 \times \begin{bmatrix} 3.3715 & 3.4610 & 3.6047 & 3.8221 & 4.1100 & 4.2684 \\ 3.4399 & 3.5317 & 3.6813 & 3.9045 & 4.1760 & 4.3166 \\ 3.5316 & 3.6327 & 3.7757 & 3.9971 & 4.2682 & 4.3978 \\ 3.6365 & 3.7207 & 3.8772 & 4.0869 & 4.3490 & 4.4987 \\ 3.7219 & 3.8000 & 3.9480 & 4.1499 & 4.4253 & 4.6193 \\ 3.7650 & 3.8431 & 3.9849 & 4.1919 & 4.4675 & 4.7109 \end{bmatrix},$$

$$[[\hat{\pi}((i,j),0)]] = \begin{bmatrix} 1.0000 & 0.9861 & 1.0000 & 1.0000 & 0.9900 & 0.0000 \\ 0.0000 & 1.0000 & 0.9900 & 1.0000 & 0.9844 & 0.0000 \\ 0.0000 & 0.9997 & 0.9910 & 0.9977 & 0.9994 & 0.0000 \\ 0.0000 & 1.0000 & 1.0000 & 1.0000 & 0.9977 & 0.0000 \\ 0.9975 & 0.9998 & 0.9884 & 0.9999 & 0.9999 & 0.0000 \\ 1.0000 & 1.0000 & 1.0000 & 0.9998 & 1.0000 & 0.0000 \end{bmatrix}.$$

Note that the convergence is slow, the price we pay for not using prior information about the transition mechanism as well as for asynchronism.

7. Further directions

We conclude by listing some promising research directions.

Average cost control: In many problems such as control of communication networks where quasi-equilibrium behaviour is desired, long-run average cost is preferred to discounted cost. It would be useful to extend the algorithm to the average cost setup, possibly using the techniques developed in Abounadi *et al* (1996a,b) for Q-learning.

Approximation based on compact representations: In addition to the approximation scheme proposed in § 5 above, Tsitsiklis & Van Roy (1996) also consider another scheme based on compact representations in the context of Q-learning. The idea is to directly approximate $V(\cdot)$ by a function belonging to a prescribed parameterized family and update the parameter in question rather than updating $V(\cdot)$ directly. In the actor–critic scheme, however, there is an additional iteration for $\pi_n(\cdot, \cdot)$. One can conceivably write a function approximation scheme for $\pi(\cdot, \cdot)$ using a parameterized family (such as a neural network – see, e.g., Santharam & Sastry 1997) and update the probabilities recursively. These possibilities need to be explored.

edback implementations: The algorithm we consider above is 'off-line', i.e., is based a simulation run rather than an actual system being controlled in real time. One can nvert it into an on-line (or feedback) adaptive control algorithm for a controlled Markov ain $X_n, n \geq 0$, by setting $Y_n = \{X_n\}$ $\forall n$ and letting $\pi_n(X_n, \cdot)$ be the actual randomised ntrol law being implemented. A natural question then is whether the scheme is asymptot-lly optimal. (For the appropriate concept of 'asymptotically optimal' in the discounted mework, see Schäl 1987). Recall that the convergence of the algorithm to desired limits quires that all state-action pairs be tried sufficiently often. For states, this may happen tomatically if suitable irreducibility conditions are met, even in the feedback case. But he $\pi_n(\cdot, \cdot)$ converge rapidly (to the desired limit or otherwise) all the state-action pairs ly not get updated frequently enough. A simple way out of this conflict is to modify the edback law to a convex combination of $\varphi_n(\cdot)$ and the uniform distribution on A so as to sure a minimum probability $\epsilon > 0$ of each $a \in A$ being picked. For $\epsilon > 0$ sufficiently all, the scheme will be nearly optimal within a prescribed tolerance. However, too small ϵ may slow down convergence. Thus there is a trade-off involved. A potentially promis-g scheme is to start with a large $\epsilon \in (0, 1/r]$ (to ensure all state-action pairs being tried quently) and then reduce it 'slowly' enough to ensure optimality. (Recall the simulated nealing algorithm for global optimization.) It is, however, a nontrivial task to capture e optimal rate of decrease of ϵ in a precise manner. These issues need further study. One uld also add that presence of interprocessor communication delays causes nontrivial mplications in the feedback case.

te of convergence: We have not provided any theoretical analysis of convergence e. Since a stochastic approximation algorithm eventually tracks the associated ODE a precise sense outlined in § 3, the convergence of its interpolated version to a given ighbourhood of the asymptotically stable limit of the ODE (assuming one exists) will osely mimic that of the ODE itself. The rate of the latter could be gauged from the apunov function approach. One must, however, invert the time-scaling $n \rightarrow t(n)$ to get e convergence behaviour of the original algorithm.

Even this may be worthless if 'eventually' is in too distant a future. There are other oblems too: The ODE captures the averaging effect of the algorithm akin to the law of ge numbers. But there can be fluctuations around the average behaviour of the 'central nit theorem' variety. For a two time scale algorithm, the time scales should be sepa-ed enough so that the slow one does not get swamped by the fluctuations of the fast e.

A related issue is the generally high variance of the stochastic approximation algorithms. additional averaging can reduce this, see, e.g., Polyak (1990). This and other issues rtaining to improving the performance of the algorithm need careful study. To cut a long ory short, the newly opened field of simulation-based algorithms for control offers many allenges both in theory and practice.

References

Abounadi J, Bertsekas D, Borkar V 1996a ODE analysis of stochastic algorithms involving sup-norm non-expansive maps (preprint)

Abounadi J, Bertsekas D, Borkar V. 1996b Q-learning algorithms for average cost problems (preprint)

Barto A, Sutton R, Anderson C 1983 Neuron-like elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybern.* 13: 835–846

Barto A, Bradtke S, Singh S 1995 Learning to act using real-time dynamic programming. *Artif. Intell. (Special Issue on Computational Theories of Interaction and Agency)* 72: 81–138

Benveniste A, Metivier M, Priouret P 1990 *Adaptive algorithms and stochastic approximations* (Berlin-Heidelberg: Springer-Verlag)

Bertsekas D, Tsitsiklis J 1989 *Parallel and distributed computation: Numerical methods* (Englewood Cliffs, NJ: Prentice Hall)

Borkar V 1994 *Asynchronous stochastic approximation. SIAM J. Control Optimization* (to appear)

Borkar V 1996 *Stochastic approximation with two time scales. Syst. Control Lett.* 29: 291–294

Brandière O, Duflo M 1996 Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. Henri Poincarè* 32: 395–427

Chazan D, Miranker W 1969 Chaotic oscillations *Linear Algebra Appl.* 2: 199–222

Hirsch M 1989 Convergent activation dynamics in continuous time networks. *Neural Networks* 2: 331–349

Keerthi S S, Ravindran B 1994 A tutorial survey of reinforcement learning. *Sādhanā* 19: 851–889

Konda V 1996 *Learning algorithms for Markov decision processes.* Master's thesis, Dept. of Computer Science and Automation, Indian Institute of Science, Bangalore

Kushner H, Clark D 1978 *Stochastic approximation for constrained and unconstrained systems* (New York: Springer-Verlag)

Neveu J 1975 *Discrete parameter martingales* (Amsterdam: North Holland)

Pemantle R 1990 Non-convergence to unstable points in urn models and stochastic approximations. *Ann. Probab.* 18: 698–712

Polyak B 1990 New method of stochastic approximation type. *Autom. Remote Control* 51: 937–946

Puterman M 1994 *Markov decision processes* (New York: John Wiley)

Santharam G, Sastry P S 1997 A reinforcement learning neural network for adaptive control of Markov chains. *IEEE Trans. Syst. Man Cybern.* 27: 588–600

Schäl M 1987 Estimation and control of discounted dynamic programming. *Stochastics* 20: 51–71

Schweitzer P, Seidman A 1985 Generalized polynomial approximations in Markovian decision processes. *J. Math. Anal. Appl.* 110: 568–582

Tsitsiklis J 1994 Asynchronous stochastic approximation and Q-learning. *Mach. Learning* 16: 185–202

Tsitsiklis J, Van Roy B 1996 Feature-based methods for large scale dynamic programming. *Mach. Learning* 22: 59–94

Walrand J 1988 *Introduction to queueing networks* (Englewood Cliffs, NJ: Prentice Hall)

Watkins C 1989 *Learning from delayed rewards.* Ph D thesis, Cambridge University, Cambridge, England

Watkins C, Dayan P 1992 Q-learning. *Mach. Learning* 8: 279–292

liams R, Baird L III 1990 A mathematical analysis of actor–critic architectures for learning optimal controls through incremental dynamic programming. In *Proc. Sixth Yale Workshop on Adaptive and Learning Systems*, New Haven, CT, pp 96–101

hizawa T 1966 *Stability theory by Liapunov's second method* (Tokyo: Mathematical Society of Japan)

An optimal fuel-injection policy for performance enhancement in internal combustion engines

V H GUPTA[1] and SHALABH BHATNAGAR[*,†,2]

[1] 9, Anand Nagar, Raipur 492 001, India
[2] Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India
[†] Present address: Institute of Systems Research, 2269 A V Williams Building, University of Maryland, College Park, Maryland, MD 20742, USA
e-mail: shalabh@isr.umd.edu

Abstract. A fuel-injection internal combustion engine system is considered, wherein supply of fuel to the engine is controlled with the twin purposes of maximizing power output and minimizing fuel wastage. The system is modelled as a controlled Markov chain and a feed-back optimal control policy is obtained for the long-run average reward optimality criterion using Markov decision theory.

Keywords. Dynamic programming/optimal control; infinite state Markov chain; long run average reward; optimal policy; fuel injection control in IC engines.

1. Introduction

Internal combustion (IC) engines are widely used in a variety of applications and almost most importantly in automobiles. A large number of such machines use the two-stroke petrol engines. One of the gravest problems with two-stroke petrol engines is the unreliable combustion, which leads to cycles (a cycle comprises one revolution of the crank and is equivalent to two strokes of the piston) that do not produce any power and also release unignited fuel which goes to the exhaust and gets wasted. Twenty-five percent of the cycles being 'powerless' is quite common even in the latest engines. Until recently, fuel was supplied to the system using a fuel-air mixer (technically known as carburettor) almost in all cases. This carburettor being in line with the air–flow of the system would supply the fuel–air mixture continuously to the combustion chamber system. However, in recent times, there is a shift to 'direct petrol injection' systems, where air supply is done independently and the fuel is directly injected into the system, once every cycle. This has rendered the

---

*Author for correspondence

system amenable to fuel injection control, since one could decide whether the fuel was to be injected or not. We refer the reader to Moskwa & Hedrics (1992) and Jones *et al* (1995), and references therein, for the latest trends in this direction.

In the following, we propose a stochastic model of the system, and also propose strategies for control. The key idea is to predict on-line, the occurrences of 'powerless' cycles stochastically, and to cut fuel at those random instants, thus minimizing fuel wastage without compromising on the power output of the engine.

2. The problem

We consider an engine with fuel injection operating in a steady state. The two binary variables that are of interest to us every ith cycle are x_i and y_i, which are defined as follows. $x_i = 0$, if fuel is not injected, and $x_i = 1$, if fuel is injected. Similarly, $y_i = 0$, if combustion does not take place, and $y_i = 1$, if combustion does take place. We note that if we define $z_i = x_i - y_i$, then z_i is also a binary variable such that $z_i = 0$, if $x_i = y_i = 1$ or $x_i = y_i = 0$, and $z_i = 1$, if $x_i = 1$ and $y_i = 0$. It is evident that the possibility $x_i = 0$ and $y_i = 1$ is impossible, since there can be no combustion without any fuel being injected. Clearly, if $z_i = 0$, then there is no wastage of fuel, and if $z_i = 1$, then fuel in that cycle is wasted. Hence, $\sum_i z_i$ is a 'measure' of the total fuel wastage. Also note that $\sum_i y_i$ is a 'measure' of the total power output. If we further denote by non-negative constants B and A, the cost of fuel per cycle and the profit or gain from power output per cycle respectively, then the problem reduces to the combined task of maximizing $\sum_i Ay_i$ and minimizing $\sum_i Bz_i$. Accordingly, we choose a cumulative reward function C, defined by $C = A \sum_i y_i - B \sum_i z_i$. Here, we also note that x_i and y_i being random variables, we cannot directly maximize C, though we could maximize the expectation of C. Noting that y_i are random variables, we also make the following assumptions.

Assumptions A. (1) $\{y_i\}$ is such that whenever $y_i = 0$, (2) and (3) below (of (A)) hold from i onward and this does not depend on the value of i.

(2) We denote by $p_1, p_2, \ldots$ the probabilities Prob$\{y_{i+1} = 1 | y_i = 0, x_{i+1} = 1\} = p_1$, and for $k > 1$,

$$\text{Prob}\{y_{i+k} = 1 | y_{i+k-1} = 1, \ldots, y_{i+1} = 1, y_i = 0, x_{i+k} = 1\} = p_k,$$

where the vertical bar denotes conditional probability.

(3) $1 > p_1 \geq p_2 \geq p_3 \geq \cdots$ etc.

We now explain the physical significance of A(1)–A(3).

A(1) implies that whenever there is a 'powerless' cycle, the whole system begins afresh and what happened before such a cycle does not affect what happens afterwards. This is quite reasonable, since otherwise one will have to assume either an infinite memory system or an arbitrarily assigned finite memory system.

A(2) means that combustion history probabilistically determines its future. One should note here that this structure must be corroborated by 'experimental observation'.

A(3) means that as the sequence of fuel utilization (fuel-non-wastage) becomes longer, the chances of encountering fuel wastage increase. This is possibly the outcome of the

pessimism that the longer things go 'right', the chances that they go 'wrong' increase! The restriction $p_1 < 1$ is mild and merely allows for consecutive 'powerless' cycles. However p_1 can be as close to 1 as we want.

In what follows, we shall formulate this problem as a controlled Markov chain and apply Markov decision theory to obtain the optimal closed loop feedback control policy.

We shall now describe the manner in which this controlled Markov chain is constructed, in lieu of (A). We use $\{0, 1\}$ as the control set with x_i as the control at the ith step. Hence, at any stage we pick $x_i = 1$ as the control if we decide to send fuel and choose $x_i = 0$ otherwise.

Let $\{X_n\}$ represent the controlled Markov chain on the state space of nonnegative integers with $\{0, 1\}$ above as the control set. Let $X_0 = j$ represent the initial state, for some integer j. Then for any integer $i \geq 1$, $X_i = j + i$ if $y_1 = y_2 = \cdots = y_i = 1$, else if $y_k = 0$ for some integer k, $X_k = 0$ as well. Hence for $i < n$, if $X_n = i$, then $X_{n-i} = 0$. Therefore any state i in the state space of the Markov chain corresponds to the ith consecutive 1 in the y_n sequence. The moment $y_n = 0$ for some n, the Markov chain jumps back to state 0. Let $p(i, j, a)$ represent the transition probability of moving from state i to state j when action or control $a \in \{0, 1\}$ is chosen. The transition probabilities then are:

$$p(i, i + 1, 1) = p_{i+1}, \quad p(i, 0, 1) = 1 - p_{i+1}, \quad \forall i \geq 0,$$

and $p(i, 0, 0) = 1, \forall i \geq 0$. Here p_i are as in (A).

Let $\hat{R}(i, a)$ represent the one-step reward random variables when the Markov chain is in state i and action a is chosen. In lieu of the cumulative reward function C defined earlier, we define $\hat{R}(i, a)$ as follows:

$$\hat{R}(i, 1) = A \quad \text{w.p. } p_{i+1},$$
$$= -B \quad \text{w.p. } 1 - p_{i+1},$$

$\forall i \geq 0$ and $\hat{R}(i, 0) = 0$ w.p. 1, $\forall i \geq 0$. Here and everywhere else w.p. stands for 'with probability'. Now let $R(i, a)$ be the one-step expected reward in state i when action a is chosen. Then,

$$R(i, 1) = A \cdot p_{i+1} - B(1 - p_{i+1}) = (A + B) \cdot p_{i+1} - B, \quad \forall i \geq 0,$$

and $R(i, 0) = 0, \forall i \geq 0$.

A policy is defined as a rule for selecting actions and can in general be randomized. Let for any policy π and initial state i,

$$\phi_\pi(i) = \liminf_{n \to \infty} \frac{E_\pi \left[\sum_{j=0}^n R(X_j, a_j) / X_0 = i \right]}{n + 1}. \tag{1}$$

Here E_π represents the expected value under policy π, $\phi_\pi(i)$ represents the average expected return per unit time when policy π is employed and initial state is i. We shall say that a policy π^* is average reward optimal if $\phi_{\pi^*}(i) = \max_\pi \phi_\pi(i), \forall i$. Our objective shall be to find π^* from the class of all policies.

Let V_α represent the value function for the α-discounted reward criterion defined as:

$$V_\alpha(i) = \max_\pi E_\pi \left[\sum_{n=0}^\infty \alpha^n R(X_n, a_n) / X_0 = i \right], \tag{2}$$

where the discount factor $\alpha \in (0, 1)$. The dynamic programming equation for the α-discounted reward tells us (Ross 1983, p. 31)

$$V_\alpha(i) = \max_a \left\{ R(i, a) + \alpha \sum_j p(i, j, a) V_\alpha(j) \right\}.$$

Since we have only two controls 0 and 1, the above equation in our case becomes

$$V_\alpha(i) = \max\{R(i, 1) + \alpha p_{i+1} V_\alpha(i + 1) + \alpha(1 - p_{i+1}) V_\alpha(0), \ \alpha V_\alpha(0)\}, \tag{3}$$

for $i \geq 0$. In what follows, we shall first show that $V_\alpha(i)$ is nonincreasing in i.

Lemma 1. Under (A), $R(i, 1)$ is nonincreasing in i.

Proof. $R(i + 1, 1) = (A + B).p_{i+2} - B \leq (A + B).p_{i+1} - B = R(i, 1)$, the inequality following because of A(3). This is true for every i. The claim follows. □

PROPOSITION 1
Under (A), $V_\alpha(i)$ is nonincreasing in i.

Proof. Consider the value iteration form of (3) (see for instance, chapter II.3 of Ross 1983 for value iteration). We have

$$V_\alpha^n(i) = \max\{R(i, 1) + \alpha p_{i+1} V_\alpha^{n-1}(i + 1)$$
$$+ \alpha(1 - p_{i+1}) V_\alpha^{n-1}(0), \ \alpha V_\alpha^{n-1}(0)\}, \tag{4}$$

$\forall i \geq 0, n \geq 1$. We shall first show by induction that $V_\alpha^n(i)$ is nonincreasing in i for every n, and then we shall show that $V_\alpha^n(i) \to V_\alpha(i)$ as $n \to \infty$, thereby implying the claim.

Let $V_\alpha^1(i) = R(i, 1)$. Then, by the previous lemma, $V_\alpha^1(i + 1) \leq V_\alpha^1(i)$. Now assume that $V_\alpha^{n-1}(j)$ decreases in j. So, as in (4), we have for $i + 1$,

$$V_\alpha^n(i + 1) = \max\{R(i + 1, 1) + \alpha p_{i+2} V_\alpha^{n-1}(i + 2)$$
$$+ \alpha(1 - p_{i+2}) V_\alpha^{n-1}(0), \ \alpha V_\alpha^{n-1}(0)\}. \tag{5}$$

We need to consider only the first terms in the arguments on the RHS of (4) and (5), since the second terms are identical. We have,

$$R(i, 1) + \alpha p_{i+1} V_\alpha^{n-1}(i + 1) + \alpha(1 - p_{i+1}) V_\alpha^{n-1}(0)$$
$$= R(i, 1) + \alpha p_{i+1}(V_\alpha^{n-1}(i + 1) - V_\alpha^{n-1}(0)) + \alpha V_\alpha^{n-1}(0). \tag{6}$$

Now since $V_\alpha^{n-1}(j)$ is nonincreasing in j, we have

$$V_\alpha^{n-1}(i + 2) \leq V_\alpha^{n-1}(i + 1) \leq \cdots \leq V_\alpha^{n-1}(0),$$

so, $V_\alpha^{n-1}(i+1) - V_\alpha^{n-1}(0) \geq V_\alpha^{n-1}(i+2) - V_\alpha^{n-1}(0)$. Now, by lemma 1 above, we have the LHS of (6) $\geq R(i+1, 1) + \alpha p_{i+2}(V_\alpha^{n-1}(i+2) - V_\alpha^{n-1}(0)) + \alpha V_\alpha^{n-1}(0)$ or, $V_\alpha^n(i) \geq V_\alpha^n(i+1), \forall i \geq 0$. Hence, $V_\alpha^n(i)$ is nonincreasing in i, $\forall n$ by induction. Since $R(i, 0) = 0$, $\forall i$, it can be easily seen by (2) and lemma 1 that $|V_\alpha(i)| \leq |R(0, 1)|/(1 - \alpha)$. Now as in Ross (1983, p. 36) $V_\alpha(i) = \lim_n V_\alpha^n(i)$, uniformly in i. Hence $V_\alpha(i)$ is nonincreasing in i. $\qquad\square$

Let $T = \min\{n > 0 | X_n = 0\}$ represent the first time the Markov chain $\{X_n\}$ hits state 0 starting from any state. Note that T is a random time. Let f_α be the α-discount optimal policy. Further, let f be the policy that chooses action 1 for all n. Now clearly, $E_{f_\alpha}[T/X_0 = i] \leq E_f[T/X_0 = i]$. Further, from (A) it is clear that $E_f[T/X_0 = i] \leq E_f[T/X_0 = 0]$, $\forall i$. Now under f with $X_0 = 0$, we have: $T = 1$ w.p. $1 - p_1$, and for $j > 1$, $T = j$ w.p. $p_1 p_2 ... p_{j-1}(1 - p_j)$. Now by A(3),

$$E_f[T/X_0 = 0] = \sum_{i=1}^\infty i \cdot p_1 \cdots p_{i-1}(1 - p_i)$$

$$\leq \sum_{i=1}^\infty i \cdot p_1^{i-1}(1 - p_i)$$

$$\leq \sum_{i=1}^\infty i \cdot p_1^{i-1}$$

$$= \frac{1}{(1 - p_1)^2} < \infty,$$

where, in the first equation above, for $i = 1$, the summand is to be interpreted as $(1 - p_1)$.

Hence $E_{f_\alpha}[T/X_0 = i] < \infty$, for all i and α. So, by theorem V.2.4 of Ross (1983), $V_\alpha(i) - V_\alpha(0)$ is uniformly bounded in i and α, as a result of which by theorem V.2.2 of Ross (1983), there exists a bounded function $h(i)$, and a constant g, such that

$$g + h(i) = \max\{R(i, 1) + p(i, i + 1, 1)h(i + 1)$$
$$+ p(i, 0, 1)h(0), \ h(0)\}, \tag{7}$$

where $g = \lim_{\alpha \to 1}(1 - \alpha)V_\alpha(0)$. Now, as a result of (7) and theorem V.2.1 of Ross (1983), there exists a stationary policy π^* such that $\phi_{\pi^*}(i) = \max_\pi \phi_\pi(i) = g, \forall i \geq 0$, and is the one which for each i, prescribes an action that maximizes the right side of (7). Again by theorem V.2.2 of Ross (1983), $\exists\{\alpha_n\}$ of discount factors such that $\alpha_n \in (0, 1), \forall n$ and $\alpha_n \uparrow 1$ as $n \to \infty$ and $h_{\alpha_n}(i) \to h(i)$ which is a bounded function, and where $h_{\alpha_n}(i) = V_{\alpha_n}(i) - V_{\alpha_n}(0)$. By proposition 1, $h_{\alpha_n}(i) \leq 0, \forall n, i$. Hence, $h(i) \leq 0 \ \forall i$. Also, $h_{\alpha_n}(0) = 0, \forall n$, and hence $h(0) = 0$. As a result, (7) becomes

$$g + h(i) = \max\{R(i, 1) + p_{i+1}.h(i + 1), 0\}. \tag{8}$$

or

$$(A + B)p_{i+1} - B + p_{i+1}h(i + 1) \leq 0,$$

or

$$p_{i+1} \leq \frac{B}{A + B + h(i + 1)}. \tag{9}$$

Let

$$\hat{i} = \min\{i \mid (9) \text{ holds}\}. \tag{10}$$

If the set in (10) is empty, we take $\hat{i} = \infty$.

The above is summarized in the following theorem, which is the main result of this section.

Theorem 1. *Under* (A), *there exists a state* $\hat{i} \leq \infty$ *given by* (10), *such that the long run average reward optimal policy selects control* 0 *when the Markov chain is in* $\hat{i}$.

4. Discussion of the results

(1) The function $h(i)$ is monotonically nonincreasing in i. This is because, by proposition 1, $V_{\alpha_n}(i + 1) \leq V_{\alpha_n}(i)$, $\forall i$, and where $\{\alpha_n\}$ is the subsequence of discount factors $\alpha \in (0, 1)$ defined earlier such that $h_{\alpha_n}(i) \to h(i)$ $\forall i$. Then, $V_{\alpha_n}(0) - V_{\alpha_n}(i + 1) \geq V_{\alpha_n}(0) - V_{\alpha_n}(i)$. Taking limits on both sides as $n \to \infty$, we obtain $h(i + 1) \leq h(i)$ $\forall i$.

(2) As a result of above, note that the RHS of (9) increases with i, whereas the LHS of (9) decreases with i and so, we would only be interested in the first i at which the RHS overshoots the LHS.

(3) $R(0, 1) > 0$ corresponds to $p_1 > B/(A + B)$. For this case, consider a policy f (say) which when starting in state i selects action 1 if $R(i, 1) > 0$, till the first k such that $R(i + k, 1) \leq 0$ (which one expects by lemma 1) and selects action 0 after that, with the same thing repeated from state 0 subsequently. If $V_{\alpha}^f(i)$ is the cumulative α-discounted reward function corresponding to this policy when starting in state i, then clearly $V_{\alpha}^f(i) \geq 0$ and so $V_{\alpha}(i) \geq V_{\alpha}^f(i) \geq 0$. Further if $R(0, 1) \leq 0$, then $R(i, 1) \leq 0$, $\forall i$, and hence it would be optimal to have a policy which selects action 0 for ever, starting in any state. It can be formally argued in this case that $V_{\alpha}(i) = 0$, $\forall i$. Hence, we conclude here that irrespective of $R(0, 1)$ being positive or negative, $V_{\alpha}(i)$ is always nonnegative for every i.

(4) From (9), define $\{\bar{p}_{\alpha_n}(i)\}$ as:

$$\bar{p}_{\alpha_n}(i) = B/(A + B + V_{\alpha_n}(i) - V_{\alpha_n}(0))$$

where $\{\alpha_n\}$ is defined as before. To check that it is a probability, we need to check that $0 \leq \bar{p}_{\alpha_n}(i) \leq 1$, $\forall i$. Now for $\bar{p}_{\alpha_n}(i) \leq 1$, we need to check if $A + V_{\alpha_n}(i) - V_{\alpha_n}(0) \geq 0$. From (3), note that $V_{\alpha_n}(i) \geq \alpha_n \cdot V_{\alpha_n}(0)$. Further as in point (3) above, $V_{\alpha_n}(0) \geq 0$, $\forall \alpha_n$. Hence,

$$A + V_{\alpha_n}(i) - V_{\alpha_n}(0) \geq A - (1 - \alpha_n)V_{\alpha_n}(0). \tag{11}$$

Now, as in § 3, $V_{\alpha_n}(0) \leq |R(0, 1)|/(1 - \alpha_n)$. If $R(0, 1) < 0$, then $V_{\alpha_n}(0) = 0$ and we are done. So, let $R(0, 1) \geq 0$. Then, since $-V_{\alpha_n}(0) \geq -R(0, 1)/(1 - \alpha_n)$, we have from (11),

$$
\begin{aligned}
A + V_{\alpha_n}(i) - V_{\alpha_n}(0) &\geq A - R(0, 1) \\
&= A - (A + B)p_1 + B \\
&= A(1 - p_1) + B(1 - p_1) > 0.
\end{aligned}
$$

Further, for $\bar{p}_{\alpha_n}(i) \geq 0$, we need denominator to be positive, and this follows immediately from above. Now, $\bar{p}_{\alpha_n}(i) \to \bar{p}(i)$, $\forall i$, as $\alpha_n \uparrow 1$, where $\bar{p}(i)$ is the RHS of (9). Hence, $0 \leq \bar{p}(i) \leq 1$, $\forall i$. It is hence, a probability.

(5) Let us consider the cases $B = 0$ and $A = 0$ separately. $B = 0$ corresponds to ignoring fuel loss which is reflected from our optimal policy, since then, $\bar{p}(\hat{i}) = 0$ and so, it is optimal to keep sending fuel for ever. This is what one expects from intuition, for maximizing power. The case $A = 0$ corresponds to ignoring power output of the engine. Note that $\bar{p}_{\alpha_n}(i) = B/(B + V_{\alpha_n}(i) - V_{\alpha_n}(0))$, in this case. However, $R(0, 1) = Bp_1 - B = B(p_1 - 1) < 0$. Hence, $V_{\alpha_n}(i) = V_{\alpha_n}(0) = 0$, as mentioned in point (3) above. So, $\bar{p}_{\alpha_n}(i) = 1$, $\forall i, n$, and hence $\bar{p}(i) = 1$, $\forall i$, again as expected, since it tells us that 'never to send fuel' policy is optimal.

(6) Finally, we look at the problem of explicitly computing g and $h(i)$. By assumption A(3), we have $0 < 1 - p_1 \leq 1 - p_2 \leq \ldots$. Consider a new process with identical state and action spaces and identical rewards, but with transition probabilities given by,

$$p(i, i + 1, 1) = \frac{p_{i+1}}{p_1}, \quad p(i, 0, 1) = 1 - \frac{p_{i+1}}{p_1} = \frac{p_1 - p_{i+1}}{p_1},$$

and $p(i, 0, 0) = 1$, $\forall i \geq 0$. Let $\bar{V}(i)$ be the p_1-discount optimal value function for this new process, where p_1 is treated as a discount factor. The analysis of (Ross 1983), pp. 98–99, carries through now, and we obtain

$$g = (1 - p_1)\bar{V}(0), \quad h(i) = \bar{V}(i) - \bar{V}(0).$$

The long-run average return optimal policy is one which selects action 0, the first time i such that $p_{i+1} \leq B/(A + B + (\bar{V}(i + 1) - \bar{V}(0)))$. This completes our discussion of the results.

5. Concluding remarks

(1) We considered the problem of controlling the supply of fuel in a fuel injection internal combustion engine system, with the twin purposes of maximizing power output and minimizing fuel wastage under very general assumptions.

(2) We formulated the problem as an infinite state controlled Markov chain on the state space of nonnegative integers using binary control variables, and applied dynamic programming techniques to obtain the long-run average reward optimal policy. The optimal policy turned out to be a threshold policy which cuts supply of fuel at a certain state $\hat{i}$ mentioned in theorem 1. As a result if $\hat{i} < \infty$, the overall Markov chain under this policy becomes finite state with state space $\{0, 1, 2, \ldots, \hat{i}\}$.

(3) We fully analysed the optimal policy. Our analysis showed that the optimal policy actually gives the threshold probability. We also considered cases in which we reduce the problem either to one of maximizing power output alone or to one in which just fuel wastage is minimized, and found that in either case the threshold policy is what one expects from intuition.

(4) Finally, we looked at the problem of explicitly computing the long-run average reward and the threshold probability. We obtained expressions for both of them in terms of p_1-discount optimal value function of a modified process, which can be easily computed. Further investigation could be carried out for the experimental verification of the algorithm.

References

Jones V K, Ault B A, Franklin G F 1995 Identification and air-fuel ratio control of a spark ignition engine. *IEEE Trans. Control. Syst. Tech.* 3: 14–21

Moskwa J J, Hedrics J K 1992 Modeling and validation of automotive engines for control algorithm development. *ASME J. Dynamic Syst. Meas. Control* 114: 278–285

Ross S M 1983 *Introduction to stochastic dynamic programming* (New York: Academic Press)

A variational formulation-based edge focussing algorithm

T J RICHARDSON[1] and S K MITTER[2]

[1] AT & T Bell Laboratories, 600-700 Mountain Av., Murray Hill, NJ 07974, USA
[2] Laboratory for Informations and Decision Systems, Massachussetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: mitter@lids.mit.edu; tjr@bell-labs.com

Abstract. Many edge detection techniques exhibit scale-dependent distortion of edges. We develop two ideas, which may also be of independent interest, to produce sharp edge localization at all scales. The first is approximation of the functional associated with the variational formulation via epi-convergence, replacing the edge set with a function. We provide a fast algorithm for minimizing the approximate functional. The second is to scale parameters and data to focus the edges. The resulting edge detector is a singular perturbation of a coupled pair of partial differential equations, yielding an elegant structure, suitable for digital or analog parallel implementation on mesh-connected arrays.

Keywords. Edge detection; parallel implementation; mesh-connected arrays.

1. Introduction

The notion of 'scale', scale of features and scale of representation, is widely held to be of fundamental importance in vision. One reason for this is that hierarchical descriptions offer potential reductions in complexity of various visual processing problems. Coarse scale segmentation of an image, for example, can be used to identify regions of interest for further processing, thereby reducing the computational load. It is important, therefore, that coarse scale descriptions retain those features of the data that are required for effective decision making. In the case of edge detection, T-junctions and corners play important roles in estimating the depth and shape of objects in a scene (Gamble & Poggio 1987). It is desirable, therefore, to accurately represent these features even at coarse scales.

Coarse scale edges have the advantage of reduced complexity; fine scale edges are more convoluted but yield more accurate and detailed information. A problem many edge-detection algorithms exhibit is systematic distortion of high curvature edges, such as T-junctions and corners; this distortion is aggravated by operating on coarse scales. This is especially true of algorithms which smooth the data in a scale-dependent way

& Hildreth 1980) and the Canny (1986) edge detector are examples of this type. More recent approaches to edge detection combat the distortion of high curvature edges by introducing interaction between image smoothing and edge placement. Examples are the Markov random field formulations (Geman & Geman 1984; Marroquin 1985; Bilbro *et al* 1992) and the related variational formulation of edge detection (Mumford & Shah 1985, 1989; Blake & Zisserman 1987). There is evidence (Blake & Zisserman 1987) that edge-detection techniques of this type do exhibit smaller localisation distortion than those of the first class. Problems remain, however; distortions still occur, and, as before, the degree of distortion depends on the scale of the edge detector. Coarse scale edges exhibit greater distortion.

Edge focussing aims at improving the localization of coarse scale edges without introducing fine scale edges. In this paper, we have derived an algorithm for performing edge focussing by starting with a variational formulation of an edge-detection problem. The resulting algorithm is described by a coupled set of nonlinear second order parabolic partial differential equations ((5)–(9) below) with explicit parameters β and c which are appropriate-adjusted (see (10)–(12)). The adjustment induces focussing of the edges. The global coarse scale nature of the edges is retained by introducing scale-stabilizing feedback mechanisms. The adjustment process commences after the nonlinear parabolic equations have nearly converged to their equilibrium. The set of equations (5)–(9) and (10)–(12) should really be viewed as an adaptive nonlinear filter which performs edge detection via focussing. Indeed, the equations are the fundamental objects in this theory and, apparently, are far more well behaved (for example, convergence to global minima) than the original variational problem.

The foundation and motivation of the adjustment of the parameters lies in certain limit theorems for the variational formulation proved by one of us (Richardson 1990, 1992). These theorems are discussed in §3. An outline of the variational formulation appears in §2. The algorithm is developed in a continuum setting in §4, and refined and discretized in §5. Simulation results can be found in §6.

The work presented in this paper is similar in spirit to that of Bergholm (1987) who focussed edges produced by the Marr–Hildreth edge detector. Since the variational formulation has better localization than the Marr–Hildreth edge detector, our approach requires less drastic adjustment of the edges.

The edge-focussing algorithm is not the only point of interest in this paper. The variational formulation is a mathematical model, not an algorithm. The primary difficulty in constructing an algorithm is appropriately representing the edges. One approach is to absorb the edges into the interaction between neighbouring pixels. This idea appears in the anisotropic diffusion approach (Nordström 1990; Perona & Malik 1990), GNC (Graduated non-convexity) type algorithms (Blake & Zisserman 1987), and in mean field annealing (Geiger & Yuille 1989; Bilbro *et al* 1992). There is another approach, which we adopt in this paper, that has been developed within the framework of approximation (of variational principles) via Γ-convergence or epi-convergence (De Giorgi & Franzoni 1979; Attouch 1984). This theory has been successfully applied to the variational formulation of edge detection by Ambrosio & Tortorelli (1990, 1992). Some computational work based on these results has appeared (March 1988, 1989, 1992). These approximations, and some further variations, are presented in §2. (The variations we indicate allow great flexibility

of the robustness of the approach.)

The approximation is achieved by reformulating the variational problem. The edge set is replaced with a function that modulates the smoothing of the image. The variational principle forces the function to have the appearance of a smoothed neighbourhood of the corresponding edge set. The degree of smoothing, i.e. the width of the effective edges, is controlled by a parameter. The 'convergence' of the approximation occurs by taking the parameter to the appropriate limit where the width of the effective edges tends to zero. One significant advantage of this reformulation is that the 'edge' function can be discretized in a straightforward way and minimization can proceed via the solution of discretized partial differential equations. In particular, the parabolic partial differential equations (5)–(9) comprise a 'gradient' descent on the approximate functional. Actually, in our implementation, we vary the step size of the 'gradient' descent and approximate a Newton-type descent (see (16)). By doing this we achieve very fast rates of convergence. Thus, we obtain a fast and elegant algorithm.

Having a separate function to represent the edges, as opposed to absorbing them into the interactions between image pixels, has some advantages. Machine Vision researchers are interested in combining various low-level vision processes, intensity edge detection and stereo depth edge detection for example, into a single operation. Having an edge representation such as the one we employ may greatly facilitate this.

It turns out that the form and the properties of the epi-convergent approximation mesh well with the parameter adjustment proposed for the edge-focussing algorithm. In particular, one can argue heuristically, and demonstrate computationally, that with 'wide' edges some edge distortions are relaxed. The price paid for this is a drop in resolution of the edges. The adjustment attempts to produce the best of both worlds, relaxing the edges initially and then sharpening them as the parameters associated with the variational formulation are adjusted for finer scales.

2. The variational formulation

Mumford & Shah (1985, 1989) suggested performing edge detection by minimizing functionals of the form

$$E(f, \Gamma) = \beta \int_\Omega (f - g)^2 \mathrm{d}\mu + \int_{\Omega - \Gamma} |\nabla f|^2 d\mu + \alpha |\Gamma|,$$

where Ω is the image domain (a rectangle), μ is Lebesgue measure, g is the observed grey level image, i.e. a real valued function, Γ denotes the set of edges, $|\Gamma|$ is the length of Γ, and β and α are real positive scalars. This approach is a modification of one due to Geman & Geman (1984) using Markov random fields, which was developed by Marroquin (1985) and by Blake & Zisserman (1987).

The three terms of E 'compete' to determine the set Γ and the function f. The first term penalizes infidelity of f to the data, while the second term forces smoothness of the approximation f, except on the edge set Γ. Thus, f is a piecewise smooth approximation to g. The third term forces some conservativeness in the use of edges by penalizing their total length. Roughly speaking, if g has a step discontinuity, approximating g with a smooth f

causes the first term to be large, tracking g more closely with a less smooth f increases the second term, and allowing Γ to approximate the support of the discontinuity reduces both the first and the second terms at some cost to the third.

This formulation was motivated in part by the desire to combine the processes of edge placement and image smoothing. Earlier edge-detection techniques such as the Marr–Hildreth edge detector, the Canny edge detector, and their variants separated these processes; the image is first smoothed, to suppress noise and control the scale, and edges are detected subsequently, as gradient maxima, for example. A consequence of this two-step approach is pronounced distortion of the edges, especially at high curvature locations. Corners tend to retract and be smoothed out; the connectedness of the edges at T-junctions is lost. The 'finger-print' images of gradient maxima of one dimensional images in scale space (Witkin 1983) are well known; the localization of edges degrades badly as scale increases. Many two-dimensional examples can be found in the literature. By introducing interaction between the edge placement and the smoothing it was expected that this effect could be abated. There is evidence, both theoretical, in one dimension (see, Blake & Zisserman 1987), and experimental, in two dimensions, that this is indeed the case.

2.1 *Approximation and computation*

To compute minimizers of E the critical question is how to represent the set Γ. A natural approach is to discretize Γ into "edge elements" and treat them combinatorially, adding or removing elements in an attempt to minimize E. Appending a stochastic component leads to the simulated annealing approach first suggested by Geman & Geman (1984). This tends to produce computationally impractical algorithms. Modifications which incorporate the edge elements into the interaction between image pixels have been proposed. One of these is based on mean field approximations of the Markov random field (Geiger & Girosi 1989; Bilbro *et al* 1992) and another, GNC (Blake & Zisserman 1987), is based on a homotopy of the interactions. Both these approaches have their strong points, and are in fact quite similar (Geiger & Yuille 1989; Bilbro *et al* 1992). A novel approach has appeared from the mathematical theory of approximation of functionals via Γ-convergence, also known as epi-convergence (De Georgi & Franzoni 1979; Attouch 1984). We will use the later terminology to avoid confusion. We will not give a general definition of epi-convergence and refer interested readers to the references cited and also to Ambrosio & Tortorelli (1990, 1992). The definition of epi-convergence is designed to allow approximation of one variational principle by another. We consider functionals of the form

$$E_c = \int_{\Omega} [\beta(f - g)^2 + \Phi(v)|\nabla f|^2$$

$$+ \alpha(c\Psi(v)|\nabla v|^2 + (1 - v)^2/4c)]d\mu. \tag{1}$$

Here $\Phi(v)$ takes the role of the Γ in E, i.e., it modulates the smoothness constraint on f. The other terms involving v force $\Phi(v)$ to simulate the effect that Γ has in E. Implicitly we have $0 \le v \le 1$. The algorithmic intention is to minimize E_c with respect to f and v. An obvious advantage the approximation offers over the original formulation is that v, since it is a function on Ω, can be discretized in a straightforward way and (local) minimizers of

t if one sets

$$\Phi(v) = v^2, \quad \text{and} \quad \Psi(v) = 1, \tag{2}$$

n E_c epi-converges to E as $c \to 0$. Some computational results for this functional have
eady appeared (March 1989) (see also Shah 1991), and a scheme similar to the one
sented here has been applied to the stereo-matching problem by March (1988).
The choice for Φ and Ψ given above may be one of the simplest possible but it is not
que. The first functional which was shown to be epi-convergent to E (by Ambrosio
Tortorelli 1990) has the form of (1) with the formal substitutions $\Phi(v) = c\Psi(v) =$
$-(1-v)^2)2c^{-(1/2)}$. When one considers algorithms based on these functionals there are
de-offs to be made between speed and performance. For example, the choice reflected in
leads to simple equations and fast computation. However, the more complicated choice
ntioned above produces sharper singularities in Φ and hence less smearing of f near
edges. With slight modifications, the proof of epi-convergence found by Ambrosio
Tortorelli (1990, 1992) can be made to go through for a large class of Ψ and Φ. In
ticular, one can choose Ψ to be any C^1 function satisfying

$$\Psi(x) > 0, \quad \text{for } x \in (0, 1],$$
$$2 \int_0^1 (1 - u)\Psi^{1/2}(u)\,du = 1.$$

te that any function satisfying the first property can be made to satisfy the second
pperty by suitable normalization. Given such a Ψ one can choose Φ to be any C^1
ction satisfying

$$\Phi(1) = 1,$$
$$\Phi(0) = 0,$$
$$\Phi(x) \in (0, 1) \quad \text{for } x \in (0, 1).$$

though the conditions given above are sufficient for the proof of epi-convergence, for
gorithms based on 'gradient' descent one should also impose the condition that Ψ be
mnotonically non-decreasing and Φ be monotonically increasing on $(0, 1)$. Furthermore,
our implementation, which is discussed in § 5, the condition $\lim_{x \to 0} \dot{\Phi}(x)/x = 0$
puld be imposed. Even more general Ψ and Φ than defined above are possible. For
ample, setting

$$\Psi(v) = \Phi(v) = \tfrac{1}{2}e^{-(1-v)^2} \tag{3}$$

o produces an epi-convergent set of functionals. Examples in the class defined above
e

$$\Phi(v) = v^{2n} \quad \text{and} \quad \Psi(v) = \frac{(m+1)^2(m+2)^2}{4}v^{2m}, \tag{4}$$

here $m \geq 0$ and $n > 0$. Equation (2) is a special case with $(n, m) = (1, 0)$.
Suppose $E(f, \Gamma) < \infty$. The proof of epi-convergence involves basically two
ps. The first is to show that for any sequence $\{f_{c_i}, v_{c_i}\}$ where $c_i \to 0$, $f_{c_i} \to f$,

and $v_{c_i} \to 1$ in an appropriate sense (not pointwise), that $\liminf_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) \geq E(f, \Gamma)$. The second is to construct a sequence such that $\limsup_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) \leq E(f, \Gamma)$. If (f, Γ) minimizes E then this second step requires constructing near minimizers of E_c. If $\liminf_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) < \infty$ then, roughly speaking, if $x \in \Gamma$ one has $\lim_{i\to\infty} \Phi(v_{c_i}(x)) = 0$. (Thus at these points we do not have $v_{c_i}(x) \to 1$, however, Γ is a set of μ measure 0.) On the other hand, the last term in (1) forces $v_c(x)$ to converge to 1 for almost all $x \in \Omega$ (in the sense of Lebesgue measure) and hence one has $\lim_{c\to 0} \Phi(v_c(x)) = 1$ for almost all $x \in \Omega$. The near minimizers of E_c are constructed by setting $\Phi(v_c(x)) \simeq 0$ on Γ and $\Phi(v_c(x)) \simeq 1$ outside some neighbourhood of Γ with a smooth transition in between. The approximations indicated here become equalities in the limit as $c \to 0$. The width of the transition depends on Ψ and on c. We give a brief heuristic description of how this occurs. In the transition region we expect f to be relatively smooth so only the terms not involving f in E_c will have a significant effect on the form of v there. In the following inequality,

$$c\Psi(v)|\nabla v|^2 + \frac{(1-v)^2}{4c} \geq \Psi^{1/2}(v)|\nabla v|(1-v),$$

the equality holds only if $|\nabla v| = \Psi^{-1/2}(v)[(1-v)/2c]$. This suggests that (in one dimension) if $u_c(t)$ satisfies $\partial u_c(t)/\partial t = [(1-u_c(t))/2c]\Psi^{-1/2}(u_c(t))$ with $\Phi(u(0)) \simeq 0$ then setting $v(x) = u_c(\text{dist}(x, \Gamma))$, for $\text{dist}(x, \Gamma) \leq \tau_c$ where $\Phi(u_c(\tau_c)) \simeq 1$ (with $u_c(\tau_c) \to 1$ as $c \to 0$), will produce near optimal transitions. This is how the near optimal v_c are constructed by Ambrosio & Tortorelli (1990, 1992). Note that assuming that $u_c(0)$ does not depend on c we obtain $u_c(t) = u_1(t/c)$. Thus the edge width is proportional to c. Let $\gamma(s) = \int_0^s (1-r)\Psi^{1/2}(r)dr$. We now compute

$$\int_0^{\tau_c} \left(c\Psi(u(t)) \left| \frac{\partial u_c(t)}{\partial t} \right|^2 + \frac{(1-u(t))^2}{4c} \right) dt$$

$$= \int_0^{\tau_c} \left(\Psi^{1/2}(u(t)) \left| \frac{\partial u_c(t)}{\partial t} \right| (1-u(t)) \right) dt$$

$$= \left| \int_0^{\tau_c} \frac{\partial}{\partial t} \gamma(u(t))dt \right| = \gamma(\tau_c) \simeq \frac{1}{2}$$

with the last approximate equality becoming equality in the limit as $c \to 0$. In the one-dimensional case we now see that the last term in (1) will contribute approximately α times the number of discontinuity points of f. In two dimensions one obtains approximately α times the length of $|\Gamma|$. Thus, we see that E_c approximates E.

Our edge-focussing algorithm will be implemented as a singular perturbation of a descent on a discrete version of E_c. We will briefly consider the continuum equivalent. Define

$$\partial_f E_c = \beta(f - g) - \nabla \cdot (\Phi(v)\nabla f), \tag{5}$$

$$\partial_v E_c = \dot{\Phi}(v)\alpha^{-1}|\nabla f|^2 - c\nabla \cdot (\Psi(v)\nabla v)$$
$$+ 2c\dot{\Psi}(v)|\nabla v|^2 + (1-v)/2c, \tag{6}$$

$$= \dot{\Phi}(v)\alpha^{-1}|\nabla f|^2 - 2c\Psi(v)\Delta v - c\dot{\Psi}(v)|\nabla v|^2 + (1-v)/2c. \tag{7}$$

The Euler–Lagrange equations for f and v are, respectively, $\partial_v E_c = 0$ and $\partial_f E_c = 0$ with Neumann boundary conditions on both v and f. Allowing f and v to depend on t we can write a 'gradient' descent on E_c in the form,

$$\frac{\partial}{\partial t} f(x, t) = -c_f \partial_f E, \tag{8}$$

$$\frac{\partial}{\partial t} v(x, t) = -c_v \partial_f E, \tag{9}$$

where c_f and c_v control the rates of descent; they would be constant for a strict gradient descent, but may not be in general. In our implementation c_v is not constant. Since the functional E_c is not jointly convex in v and f we do not expect to always reach a global minimum by a descent method. Thus the solution obtained will depend on the initial conditions and also on the parameters c_f and c_v.

Equation (8) strongly resembles the anisotropic diffusion scheme of Perona & Malik (1990) and, even more strongly, the 'biased' anisotropic diffusion scheme of Nordström (1990). Perona & Malik (1990) begin by considering the standard diffusion equation,

$$\frac{\partial}{\partial t} f(x, t) = \Delta_x f(x, t), \quad f(x, 0) = g(x),$$

which produces the 'scale-space' smoothings of g, parametrized by t. For fixed t one obtains $f(x, t)$ by convolving $g(x)$ with a Gaussian kernel whose variance is linear in t. Perona & Malik (1990) suggested controlling the diffusion coefficient to prevent smoothing across edges. Thus they were led to consider equations of the form

$$\frac{\partial}{\partial t} f(x, t) = \nabla_x \cdot (h(|\nabla_x f(x, t)|)\nabla_x f(x, t)), \quad f(x, 0) = g(x).$$

They experimented with $h(s) = J/[1 + (s/K)^2]$ and $h(s) = e^{-(s/K)^2}$, where J and K are constants. Equation (8) resembles this equation in that it is a diffusion with controlled conductivity. The control of the conductivity depends on $|\nabla f|$ indirectly through (9). The term $\beta(f - g)$ in (8) stabilizes the solution at some particular scale. Such a term also appears in the 'biased' anisotropic diffusion scheme studied by Nordström (1990). Perona & Malik (1990) analyse their scheme to show that the maximum principle holds, i.e. that the solution's extrema never exceed those of the original image. They argue that this implies that no new 'features' (blobs) are introduced into the solution. Here, as in Nordström (1990), this property is a trivial consequence of the formulation. The functionals E_c would *increase* if such new features appeared. (Truncating such a new feature would decrease E_c.) An advantage of the scheme presented here is that it yields an explicit representation of the edges (via the function $\Phi(v)$). The resulting system of equations admits a particularly simple implementation in digital mesh connected parallel machines with simple processors or, potentially, in an analog network, such as discussed by Harris *et al* (1989).

3. Scale, noise, and accuracy

Witkin (1983) introduced the idea of a scale-space representation of an image – smoothing the image and identifying edges on all scales. One of the problems arising from the scale-space concept was the correspondence problem: Which of the fine scale edges correspond to
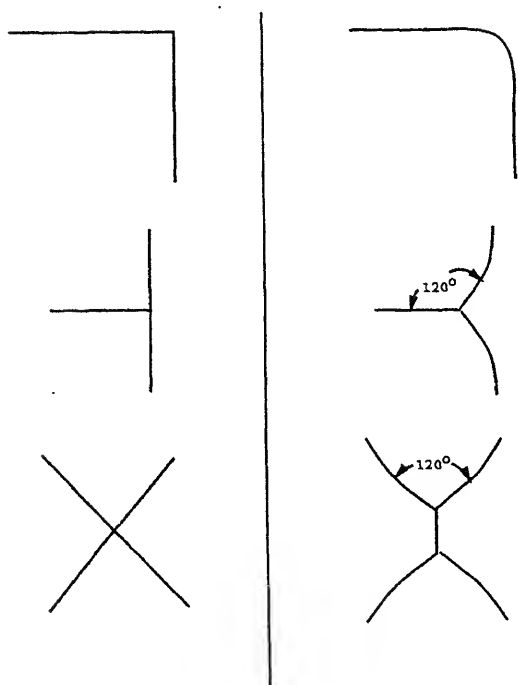
Non-minimal Geometries   Corresponding Minimal Geometries   **Figure 1.**   Calculus of variations results.

the coarse scale ones? The problem is aggravated by the fact that the distortion of the edges, mentioned earlier, depends on scale. Typically 'coarse scale' implies more smoothing and, hence, more distortion. This is undesirable in many situations because salient features, corners and T-junctions for example, tend to be obscured. The correspondence problem is therefore of great importance. The edge-focussing algorithm of Bergholm (1987) addresses the correspondence problem by taking small steps in scale. We address it this way too, in effect, but also by using a better underlying edge detection algorithm.

Since the variational approach combats the distortion caused by smoothing one hopes that the correspondence problem will be alleviated by using it. Although this appears to be the case, problems remain; there still are distortions, depending on scale, and the model intrinsically restricts the geometry of possible edge sets in an unnatural way. The analysis of Mumford & Shah (1985) showed that edge sets produced by the variational approach have the following properties, which are illustrated in figure 1.

If Γ is composed of $C^{1,1}$ arcs then

- at the most three arcs can meet at a single point and they do so at $120°$,

- they meet $\partial\Omega$ only at an angle of $90°$,

- it never occurs that exactly two arcs meet at a point (other than the degenerate case of two arcs meeting at $180°$), i.e., there are no corners.

These results derive from the fact that the term $|\Gamma|$ in E locally dominates the behaviour of singularities in Γ. Hence the types of singularities observed are identical to those of

imal surfaces. Of course 'real' edges are not restricted to these geometries. The de-
dence on scale derives from the interaction between the singularities of f and those of
Roughly speaking, smaller values of β produce greater distortion. This becomes more
ar in light of the theorems quoted below and the analysis presented by Mumford & Shah
89).

n view of the foregoing it is natural to ask what are the noise/accuracy/scale tradeoffs
l, more generally, whether scale-dependent distortions are necessary. Is it necessary that
mechanisms used to combat noise, i.e. smoothing, be tied simultaneously to scale and
uracy? We propose that the accuracy of edge localization need not be limited by scale.
e central ideas of the method we will develop can be gleaned from some asymptotic
orems proved by one of us (Richardson 1990, 1992) concerning minimizers of E. We
l now present a slightly simplified statement of those theorems.

To measure the disparity between one edge set and another we introduce the Hausdorff
tric. For $A \subset \mathbb{R}^2$ the ϵ-neighbourhood of A will be denoted by $[A]_\epsilon$ and is defined by
$|_\epsilon = \{x \in \mathbb{R}^2 : \inf_{y \in A} ||x - y|| < \epsilon\}$ where $|| \cdot ||$ denotes the Euclidean norm. Denoted
$d_H(\cdot, \cdot)$, the Hausdorff metric is evaluated by

$$d_H(A, B) = \inf\{\epsilon : A \subset [B]_\epsilon \text{ and } B \subset [A]_\epsilon\}.$$

mentary analysis shows that d_H is a metric on the space of non-empty compact sets
R^2.

Suppose for the moment that we have ideal data: g is a piecewise smooth function.
make this clear we denote it by g_I. We assume that there exists a set Γ_g, a union
curves, satisfying length $(\Gamma_g) < \infty$ such that g_I is discontinuous on Γ_g and smooth
ewhere. More precisely we require that $\int_{\Omega - \Gamma_g} |\nabla g_I|^2 d\mu < \infty$, that there exists a
nstant L such that g_I, restricted to any straight line segment lying in $\Omega - \Gamma_g$, is a Lipschitz
iction with Lipschitz constant L, and that g actually has a discontinuity everywhere on
except, possibly, for a set having zero total length. Under these conditions the following
lds.

eorem 1. *For any fixed $\alpha > 0$ and $\epsilon > 0$ there exists $\beta^\star < \infty$ such that if $\beta \geq \beta^\star$ and
is minimal for E then*

$$d_H(\Gamma_g, \Gamma_\beta) < \epsilon.$$

e thrust of this theorem is that as $\beta \to \infty$ the set Γ_β can assume arbitrary geometries
l will coincide with the discontinuity set of g_I. The theorem can be interpreted as
asymptotic fidelity result for the variational approach. It implies that the distortions
ulting from *ad hoc* functions of this type are local, small-scale effects.
From a practical point of view the result is lacking because $\beta \to \infty$ forces f to match g_I
actly; noise in g_I will result in the appearance of many spurious boundaries. However,
e theorem can be extended to incorporate noise and smearing effects. Roughly speaking,
the admissible noise magnitude scales as $o(\beta^{-1/2})$ and the admissible smearing acts
er a radius of $o(\beta^{-1})$ then their presence can be tolerated and the theorem still holds. To
ve a more precise statement of this, let h_r and h_λ be any positive real valued functions

satisfying

$$\lim_{\beta \to \infty} \beta h_r(\beta) = 0,$$

$$\lim_{\beta \to \infty} \beta^{\frac{1}{2}} h_\lambda(\beta) = 0,$$

and let $\Psi(\beta)$ be the set of all functions g which can be written as

$$g = S_r(g_I) + \lambda \omega,$$

where $0 \leq \lambda \leq h_\lambda(\beta)$ is a scalar, $\omega \in L^\infty(\Omega)$ with $|\omega|_\infty \leq 1$, S_r is any smearing operator satisfying

$$S_r f(x) \in \left[\min_{y \in B_r(x)} f(y), \max_{y \in B_r(x)} f(y) \right],$$

where $B_r(x)$ is the disc of radius r centred at x with $0 \leq r \leq h_r(\beta)$, and g_I is the ideal data described earlier. We then have the following.

Theorem 2. *For any fixed $\alpha > 0$ and $\epsilon > 0$ there exists $\beta^* < \infty$ such that if $\beta \geq \beta^*$ and Γ_β is minimal for E for some $g \in \Psi(\beta)$ then*

$$d_H(\Gamma_g, \Gamma_\beta) < \epsilon.$$

The proofs are beyond the scope of this paper (they require a lengthy function analytic development) and have appeared in Richardson (1990, 1992). The relevant mathematical framework is outlined in Ambrosio (1989). This theorem indicates how noise and localization defects should scale with the parameter β to maintain fidelity.

Remark. The requirement $\lim_{\beta \to \infty} \beta^{1/2} h_\lambda(\beta) = 0$ seems to be necessary in general. There are certain images which asymptotically produce spurious edges for the piecewise constant version of the variational formulation if this condition is violated (Richardson 1990, 1992).

Our immediate goal, fulfilled in the next section, is to describe a heuristic which mimics the theorems to produce accurate localization of the edges in a manner independent of scale. Two problems immediately suggest themselves. First, a real image has fixed noise which cannot be scaled since it cannot be identified, and secondly, smearing is fixed and cannot (in general) be removed. The intrinsic noise and smearing of an image limits the recoverable accuracy of the edge locations. We are not proposing a scheme to eliminate that. Edge detection, ostensibly, will be performed on different scales. As we have indicated, operating on coarse scales tends to introduce distortions above and beyond those inherent in the signal. Our goal is to eliminate those distortions and recover, even on coarse scales, the same accuracy of localization usually achieved only at fine scales.

Since we intend to use an approximation to the variational formulation, it is prudent to consider whether the approximation deviates in a significant way from the original formulation with regard to distortion of edges. Although analysis is prohibitively difficult, we expect (and simulations have borne this out) that the spreading of the edges in the epi-convergent approximation actually ameliorates some of the geometric distortion. We recall

that the primary reason for the geometric distortion is that the term $|\Gamma|$ in E determines the structure of the singularities. Roughly speaking, this arises because the length term is one-dimensional and scales linearly while the other terms are two-dimensional integrals and hence scale quadratically in the size of the domain. (Actually this is not precisely true because singularities arise in f, but the dominance of the length term still occurs at singularities in Γ.) When the edges are smeared and length is replaced by a two-dimensional integral the concentration of cost in the length term is alleviated and, hence, we expect the distortion to be relaxed. The price paid for this is the lack of resolution of the edges. The algorithm begins with thick edges, thus relaxing distortion, but ends by sharpening the edges while scaling the parameters in accordance with theorem 2. Thus the edges are focussed as resolution increases.

4.   Edge focussing via scaling

In this section we describe in detail the modifications to the descent equations, (8) and (9), which we introduce to focus edges. The essential idea is to satisfy the conditions of theorem 2 by smoothing g in a controlled way while allowing $\beta \to \infty$. Simultaneously we allow $c \to 0$ to sharpen the edges. We draw an analogy between the width of the edges and the diameter of the smearing operators appearing in theorem 2. Thus the rates of change of β and c are coupled.

We consider introducing dynamics into quantities β, g, and c, which in a standard minimization of E_c would be held fixed. These dynamics are intended to come into effect only after the basic descent equations (4) and (5) have essentially converged. Thus we assume $g(x, 0)$ is the initial data and $f(x, 0)$ and $v(x, 0)$ satisfy their respective Euler–Lagrange equations with $\beta = \beta(0)$ and $g(x) = g(x, 0)$. This implies the presence of a nominal set of edges, i.e. a function $v(x, 0)$. We will be guided by the heuristic that the subsequent focussing should only focus the edges already found and not introduce new ones. We make the following correspondences with the quantities which appear in theorem 2,

$$g(x, t) \longleftrightarrow S_{r(t)}(g_I) + \lambda(t)w(x, t),$$
$$f(x, t) \longleftrightarrow S_{r(t)}(g_I),$$
$$\text{where} \quad r(t) = Kc(t) \text{ for some constant } K,$$
$$|\lambda(t)|_\infty = |g(x, t) - f(x, t)|_\infty,$$
$$\text{and} \quad |w(x, t)|_\infty = 1.$$

We will discuss the choice of K and the meaning of the correspondences shortly. Consider first the following equations,

$$\frac{\partial}{\partial t} g(x, y, t) = \epsilon(f(x, y, t) - g(x, y, t)),$$

$$\frac{\partial}{\partial t} \beta(t) = \epsilon\beta(t),$$

where ϵ is a small positive constant included to reflect the fact that these equations are perturbations of (8) and (9). We observe that we obtain a solution to these equations such

$$v(x, t) = v(x, 0),$$
$$f(x, t) = f(x, 0),$$
$$\beta(t) = \beta(0)\, e^{\epsilon t},$$
$$g(x, t) = g(x, 0)\, e^{-\epsilon t} + f(x, 0)(1 - e^{-\epsilon t}).$$

In fact, if $(f(x, 0), v(x, 0))$ minimizes E_c with data $g(x, 0)$ and parameter $\beta(0)$ then it is easy to see that $(f(x, t), v(x, t))$ minimizes E_c with data $g(x, t)$ and parameter $\beta(t)$. These equations show that with this scaling we have $\beta(t) \to \infty$ while the minimal solutions $v(x, t)$ and $f(x, t)$ remain fixed. Interpreted in view of theorem 2 this means that g_I corresponds to $f(x, 0)$ (i.e. $K = 0$). Note that according to our correspondences we have $\lambda(t) \propto \beta^{-1}(t)$, so the scaling conditions of theorem 2 are satisfied. We will now alter these equations slightly. First, we will introduce some dynamics into c to sharpen the edges. Second, we suppress the smoothing of g in a neighbourhood of the edges which shrinks with time; this is to permit focussing of the edges. Thus we consider

$$\frac{\partial}{\partial t} g(x, t) = \epsilon \rho(v(x, t))(f(x, t) - g(x, t)), \tag{10}$$

$$\frac{\partial}{\partial t} \beta(t) = \epsilon \beta(t), \tag{11}$$

$$\frac{\partial}{\partial t} c(t) = -\epsilon c(t), \tag{12}$$

where $\rho(v(x, t))$ should be approximately zero inside some neighbourhood of the edges and approximately one outside some larger neighbourhood. Furthermore, the width of the larger neighbourhood should shrink as $\beta^{-1}(t)$. A simple and reasonable choice, for example, is $\rho = \Phi$ since in this case the neighbourhood width is proportional to $c(t)$, which in turn is proportional to $\beta^{-1}(t)$. The algorithm takes the form of (8) and (9) until a local minima is reached, and subsequently (10)–(12) come into effect. Assuming $g(x, t)$ and $f(x, t)$ converge, we interpret the limit $g(x, \infty) = f(x, \infty)$ as corresponding to g_I and the set $S_g = \{x : \Phi(v(x, \infty)) \simeq 0\}$ as corresponding to Γ_g. The quantity K does not appear in our equations and is meant only to facilitate the correspondences; it should be interpreted as being sufficiently large so that the set $\{x : \rho(v(x, t)) < 1 - \delta\}$ for some small positive constant δ is contained in $[S_g]_{r(t)}$. We expect that $f(x, t) \simeq f(x, 0)$ for all $x \notin [S_g]_{r(0)}$. Given this, the correspondence of $f(x, t)$ with $\Phi_{r(t)}(g_I)$ is consistent with the noise scaling of theorem 2.

In general the choice of ρ is a delicate issue. If g is noisy it may be desirable to allow more smoothing near the edges. The price for this is admitting potential distortion into the edges. In this situation a better smoothing mechanism might be a directionally controlled diffusion of g, suppressing diffusion across edges but enhancing diffusion parallel to the edge. This can easily be implemented within the framework developed here since $\nabla \Phi(v(x))$ will be perpendicular to an edge in a neighbourhood of that edge. (We have not experimented with this alternative.) Even when noise is not an issue ρ must be chosen carefully. We contend that the ideal choice should allow for edge correction and adequate smoothing without the introduction of edges from finer scales.

ne conditions, is to allow β to depend on x and to scale it only around the edges. For
mple, one could replace (11) with $(\partial/\partial t)\beta(x, t) = \epsilon(1 - \rho(v(x, t)))\beta(x, t)$ and elimi-
e the smoothing of g, (10). We have experimented with this variation in our simulations
l the results are similar to those presented here.

Discretization and parameter choices

this section we address some of the issues which arise as a consequence of dis-
tization and further refine the algorithm. The lattice spacing (we will consider only
lare lattices) can be absorbed, via scaling, into the other parameters, β, c, and α.
propriate step sizes for the discrete versions of the descent algorithm must be
ind, and the relative rates of the gradient descent and the scaling dynamics must be
cided.
For the simulations presented in this paper f, g and v are discretized in a manner
scribed below. Discrete versions of f and g are defined on a rectangular subset of a
lare lattice while the discrete version of v is defined on a similar subset of a square
tice which is twice as dense and rotated by $45°$. This is not necessary, but it facilitates
e discrete implementation. We define the following subset of $\mathbb{Z}^2 \subset \mathbb{R}^2$,

$$\mathcal{L}_f = \{(i, j) : i = 1, \ldots, N, \ j = 1, \ldots, M\}.$$

e assume that g is defined on $\mathcal{L}_f$. The nearest neighbours of $x \in \mathcal{L}_f$ are defined by

$$\mathcal{N}_f(x) = \{x' \in \mathbb{Z}^2 : |x' - x| = 1\}.$$

nilarly, we define

$$\mathcal{L}_v = \{(x + x')/2 : x \in \mathcal{L}_f, x' \in \mathcal{N}_f(x) \cap \mathcal{L}_f\},$$

ich will support v, and the nearest neighbours

$$\mathcal{N}_v(y) = \{y' = (x + x')/2 : x \in \mathcal{L}_f, x' \in \mathcal{N}_f(x), |y - y'| = 1/\sqrt{2}\}.$$

r discretization we can take the approach of discretizing E_c and then deriving discrete
ler-Lagrange equations, or we can discretize the Euler–Lagrange equations directly.
e consider the first approach first. A discrete version of E_c with lattice spacing δ is the
lowing,

$$E_d = \beta \sum_{x \in \mathcal{L}_f} \left(\delta^2 (f(x) - g(x))^2 + \frac{1}{2} \sum_{x' \in \mathcal{N}_f(x)} \Phi((x + x')/2)(f(x) - f(x'))^2 \right)$$

$$+ \alpha \sum_{y \in \mathcal{L}_v} \left(\frac{c}{4} \sum_{y' \in \mathcal{N}_v(y)} (\Psi(y) + \Psi(y'))(v(y) - v(y'))^2 + \frac{\delta^2}{8c}(1 - v(y))^2 \right),$$

ere if $x \in \mathcal{L}_f$ and $x' \in \mathcal{N}_f(x) \backslash \mathcal{L}_f$ then we impose the condition $f(x') = f(x)$ (this
fines $f(x')$ uniquely), and if $y \in \mathcal{L}_v$ and $y' \in \mathcal{N}_v(y) \backslash \mathcal{L}_v$ then we impose the condition

$v(y') = v(y'')$, where $y'' \in \mathcal{N}_v(y)$ is the unique point satisfying $|y' - y''| = 1$. This is to ensure Neumann type boundary conditions in the descent equations given below. By making the substitutions

$$\beta \to \delta^{-2}\beta, \quad c \to \delta^2 c, \quad \alpha \to \delta^{-2}\alpha,$$

we find that without loss of generality we can set $\delta = 1$, which we do henceforth. A discrete form of the Euler–Lagrange equations can now be found directly by differentiating E_d with respect to $v(y)$ and $f(x)$. To simplify the notation we will write $\Psi(y)$, $\Phi(y)$ instead of $\Psi(v(y))$, $\Phi(v(y))$. For each $y \in \mathcal{L}_v$ the pair x, x' such that $y = (x + x')/2$ and $x \in \mathcal{L}_f$, $x' \in \mathcal{N}_f(x) \cap \mathcal{L}_v$ is uniquely determined. Thus, for each such y we can set $df(y) = (f(x) - f(x'))^2$. The derivatives of Ψ and Φ will be denoted $\dot{\Psi}$ and $\dot{\Phi}$ respectively. For each $x \in \mathcal{L}_f$ and $y \in \mathcal{L}_v$ we define

$$\partial_x E_d = \beta(f(x) - g(x)) + \sum_{x' \in \mathcal{N}_f(x)} \Phi((x + x')/2)(f(x) - f(x')) \tag{13}$$

$$\partial_y E_d = \alpha^{-1}\dot{\Phi}(y)df(y) - \frac{1 - v(y)}{4c} + c \sum_{y' \in \mathcal{N}_v(y)} (v(y) - v(y')),$$
$$\times \left(\Psi(y) + \Psi(y') + \frac{1}{2}\dot{\Psi}(y)(v(y) - v(y')) \right), \tag{14}$$

which are proportional to $[\partial/\partial f(x)]E_d$ and $[\partial/\partial v(y)]E_d$ respectively. These equations are discrete analogs of (5) and (6) respectively. (The constants in (14) are slightly different from those in (6) because the lattice spacing of $\mathcal{L}_v$ is $1/\sqrt{2}$.) Now we consider discretizing the Euler–Lagrange equations directly. A discrete version of (7) is

$$\partial_y E_d = \alpha^{-1}\dot{\Phi}(y)df(y) - \frac{1 - v(y)}{4c} + c \sum_{y' \in \mathcal{N}_v(y)} (v(y) - v(y'))$$
$$\times \left(2\Psi(y) - \frac{1}{2}\dot{\Psi}(y)(v(y) - v(y')) \right). \tag{15}$$

A third alternative is to average (6) and (7) and discretize, or equivalently, to average (14) and (15). If we do so, we obtain

$$\partial_y E_d = \alpha^{-1}\dot{\Phi}(y)df(y) - \frac{1 - v(y)}{4c}$$
$$+ c \sum_{y' \in \mathcal{N}_v(y)} (v(y) - v(y')) \left(\frac{3}{2}\Psi(y) + \frac{1}{2}\Psi(y') \right).$$

From the point of view of implementation this form is much more attractive since it simplifies computation. Finally, one could approximate $3/2\Psi(y) + 1/2\Psi(y') \simeq 2\Psi(y)$ to obtain an even simpler form. Although this deviates from the theory, we use it in our implementation.

Allowing all quantities to depend on t, our basic descent equations take the form,

$$f(x, t + 1) - f(x, t) = -c_f \partial_x E_d,$$
$$v(x, t + 1) - v(x, t) = -c_v \partial_v E_d,$$
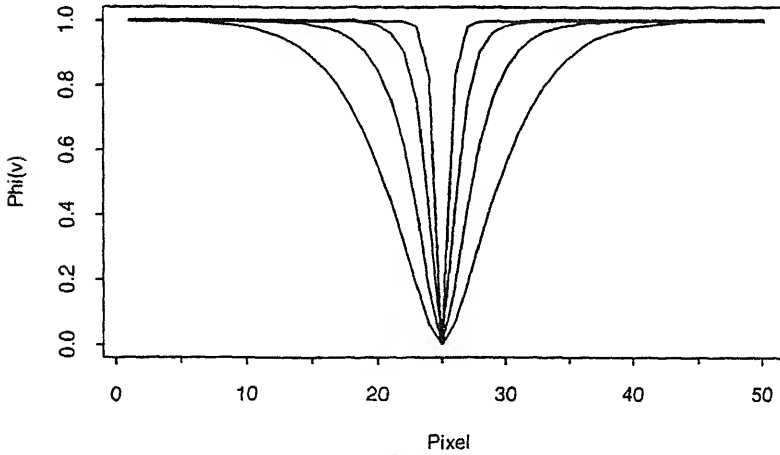
Figure 2. Graphs of minimal $\Phi(v(25, j + 0.5))$ for step edge for $c = 2.0, 1.0, 0.5, 0.2$. In each case $\alpha = 0.05, \beta = 2.0,$ and Φ and Ψ were chosen as in (2).

ere on the right hand side all quantities are evaluated at time t. We now address the estion of choosing c_f and c_v. A standard gradient descent would have both c_f and constant. If we try to set c_v constant then it must be chosen small since $(f(x) - x'))^2/\alpha$ can be quite large and hence convergence will be slow. A computationally effient choice that gives much faster convergence is to approximate a Newton-type descent. If set

$$c_v(y) = \frac{1}{2}\left(\frac{\dot{\Phi}(y)}{\alpha v(y)}\mathrm{d}f(y) + \frac{1}{4c} + 8c\Psi(y)\right)^{-1},$$
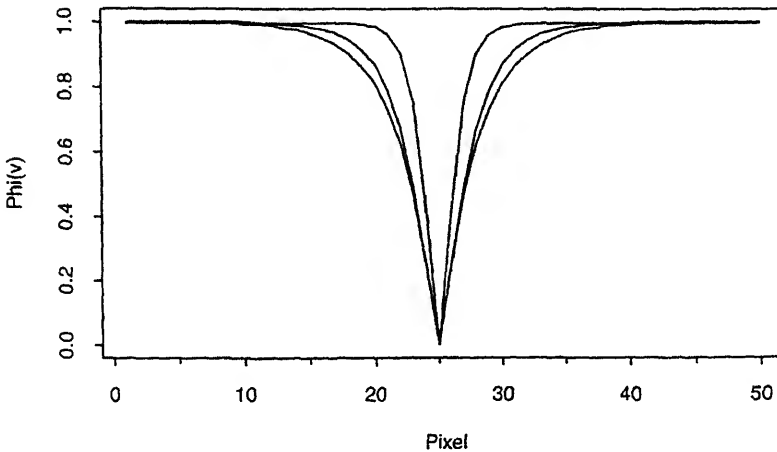
Figure 3. Graphs of minimal $\Phi(v(25, j + 0.5))$ for step edge for $\alpha = 0.05, \beta = 2.0,$ and $c = 0.5$ where Φ and Ψ are given by (4) with (n, m) given by $(2, 0)$ $(2, 1/2)$, and $(2, 1)$ respectively where increasing m increases the spread of the function.

above, then we obtain

$$v(y, t+1) = \frac{1}{2}\left(v(y,t) + \frac{1/4c + 2c\Psi(y)\sum_{y'\in\mathcal{N}_v(y)} v(y')}{\Phi(y)df(y)/(\alpha v(y)) + 1/4c + 8c\Psi(y)}\right). \quad (16)$$

Our implementation employs this form of update. It has particularly attractive properties. Note, in particular, that $v(y, t+1)$ is an average of $v(y, t)$ and a well-behaved quantity which lies between 0 and 1. Setting c_f constant is much less problematic, and for our simulations we have set it to $(2\beta + 8)^{-1}$ since this gives a good rate of convergence without allowing overshoot in f. The initialization of f and v will affect which local minimum is reached by the initial gradient descent. We expect this will have little effect on the edge-focussing part of the algorithm. In our simulations we have set $f(0) = g(0)$ and $v(0) = 1$. With these choices we observe that the basic descent on f and v converges in about 30 iterations for the range of parameters we have experimented with. (Larger values of c and smaller values of β will reduce the rate of convergence.)

We now consider the discrete scaling dynamics. We recall that these equations come into effect only after the gradient descent has nearly converged. The following are discrete analogs of (10)–(12),

$$g(x, t+1) = g(x, t) + \epsilon\rho(x, t)(g(x, t) - f(x, t)),$$
$$\beta(t+1) = (1 - \epsilon)^{-1}\beta(t),$$
$$c(t+1) = (1 - \epsilon)c(t),$$

where ρ and ϵ are to be specified in each case. Approximations that save computation can be made here. For example one could compute this update once every n time steps instead of every time step and increase ϵ appropriately. Also, since $f(x, t)$ remains essentially fixed outside of some neighbourhood of the edges, one could restrict the update on β to some neighbourhood of the edges and completely drop the update on g producing essentially the same effect, as mentioned earlier. This would yield substantial time savings in serial implementations. Our implementation employs the equations given above.

If $c(t)$ becomes too small then the discrete approximation to E_c will break down. The value of $c(t)$ can be used as a stopping criteria. For the choices of Ψ and Φ used in this paper we allow $c(t)$ to become small enough so that the effective edge width is one pixel. (Effective edge width can be defined as the width of the set $\{\Phi(y) < 1/2\}$ for example.)

6. Simulation results

For all of our simulation results we have scaled g so that $g(x) \in [0, 1]$. In particular, solid white is 1 and solid black is 0. In the two-dimensional plots and in the images we plotted $\Phi(v)$ on the same mesh as f, i.e. the plotting mesh corresponds to $\mathcal{L}_f$. For each $x \in \mathcal{L}_f$ we plot the minimum value of $\Phi(v)$ among the four nearest neighbours. (Note that an edge 'set' could be defined as $\{y \in \mathcal{L}_v : \Phi(v(y)) < 1/2.)$

The first two simulations illustrate basic properties of the epi-convergent approximation. Here g is defined on a 50×50 square mesh with $g(i, j) = 0$ for $j \leq 25$ and 1 otherwise. Figure 2 illustrates dependence of the edge smearing on c. We plot $\Phi(v(25, j+0.5))$, $j =$
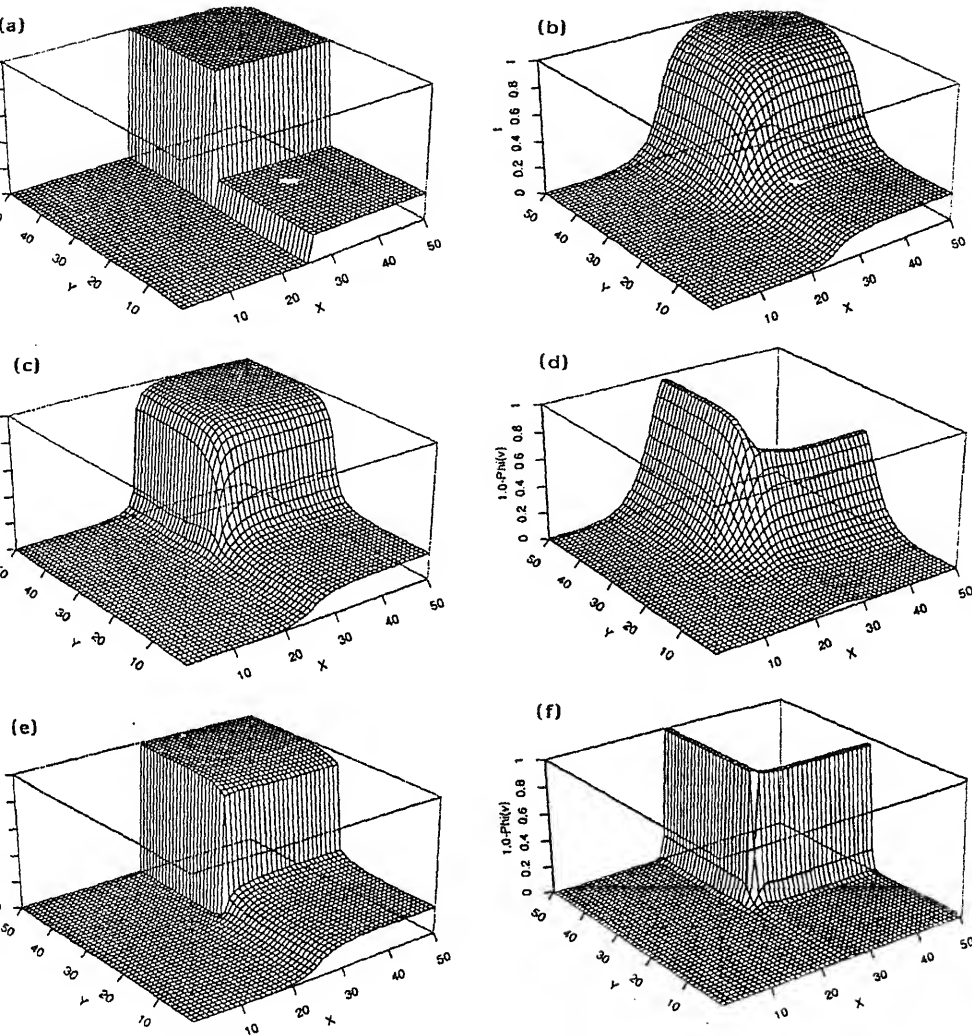
Figure 4. (a) Synthetic T-junction/corner data. (b) Minimal f with $v = 1$ with $\beta = 0.1$. (c) Minimal f with $\beta = 0.1$, $\beta = 0.1$, and $c = 2.0$ (d) Minimal $\Phi(v)$ with $\beta = 0.1$, $\alpha = 0.1$, and $c = 2.0$ (e) Minimal f after edge focussing with final values of $\beta = 1.0$, $\alpha = 0.1$, and $c = 0.2$. (f) Minimal $\Phi(v)$ after edge focussing with final values of $\beta = 1.0$, $\alpha = 0.1$, and $c = 0.2$.

.., 50 for $c = 2.0, 1.0, 0.5, 0.2$ after convergence of the descent equations (without
ling). In each case $\alpha = 0.05$, $\beta = 2.0$, and Φ and Ψ were chosen according to (2).
ure 3 is similar except that we have fixed $c = 0.5$ and varied Φ and Ψ. They are given
(4) with $(n, m) = (1, 0), (1, 1/2). (1, 1)$ respectively (increasing m tends to increase
width of the edge). This is to illustrate how the shape of the edges can be changed by
ring Φ and Ψ.
o demonstrate the behaviour of the edge focussing algorithm on high curvature edges
have simulated the algorithm on the data presented in figures 4 a–f. The functions Φ
Ψ are as in (2) and we set $\rho(v) = \Phi(v)$. We have carefully chosen the parameters
make the detection of the corner marginal. The initial values are $\beta(0) = 0.1$, $\alpha = 0.1$,
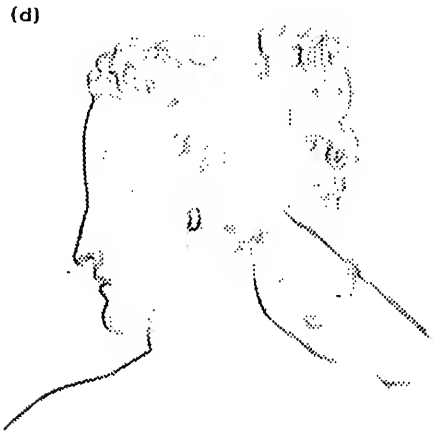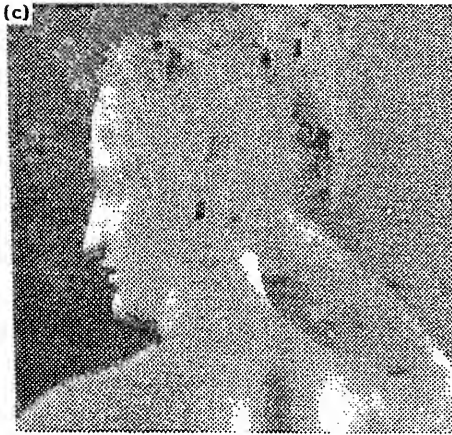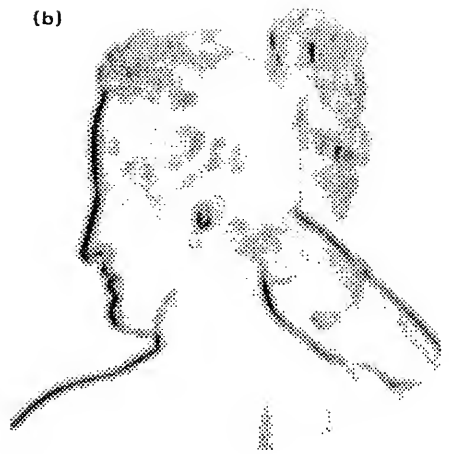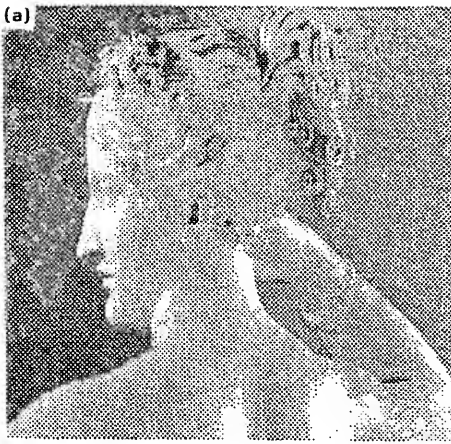
Figure 5. (a–d) (Caption on facing page.)

and $c(0) = 2.0$. The final values are $\beta(T) = 1.0$ and $c(T) = 0.2$ (where T is the time of termination), and ϵ was chosen so that 200 iterations with scaling are required. Figure 4a is the data set g. Figure 4b plots the optimal f when $\alpha = \infty$, i.e. no edges are allowed. This is to indicate the degree of smoothing associated with this value of β. Figure 4c shows $f(0)$, i.e. the f obtained after allowing a descent to converge with the parameters held fixed at their initial values. Figure 4d shows $\Phi(v(0))$. We have chosen the parameters so that the three edges in the image are detected to different degrees. Figures 4e and 4f are $f(T)$ and $\Phi(v(T))$ respectively. Note that the two larger edges and the corner they form is unambiguously detected while the smallest edge has been smooth out. The slight smoothing visible across the second largest edge is due to slight smoothing feedback at that edge. The function $g(T)$ is not distinguishable from $f(T)$ so we have not included it.

Figures 5 and 6 demonstrate the algorithm on 'real' images. Figure 5 is 480×512 pixels and figure 6 is 512×512 pixels. In general ϵ is chosen so that 200 iterations with scaling are required. The data are in figure (a) in each case. Each image has been processed for two different sets of parameters to indicate the stability of the edges under a change in
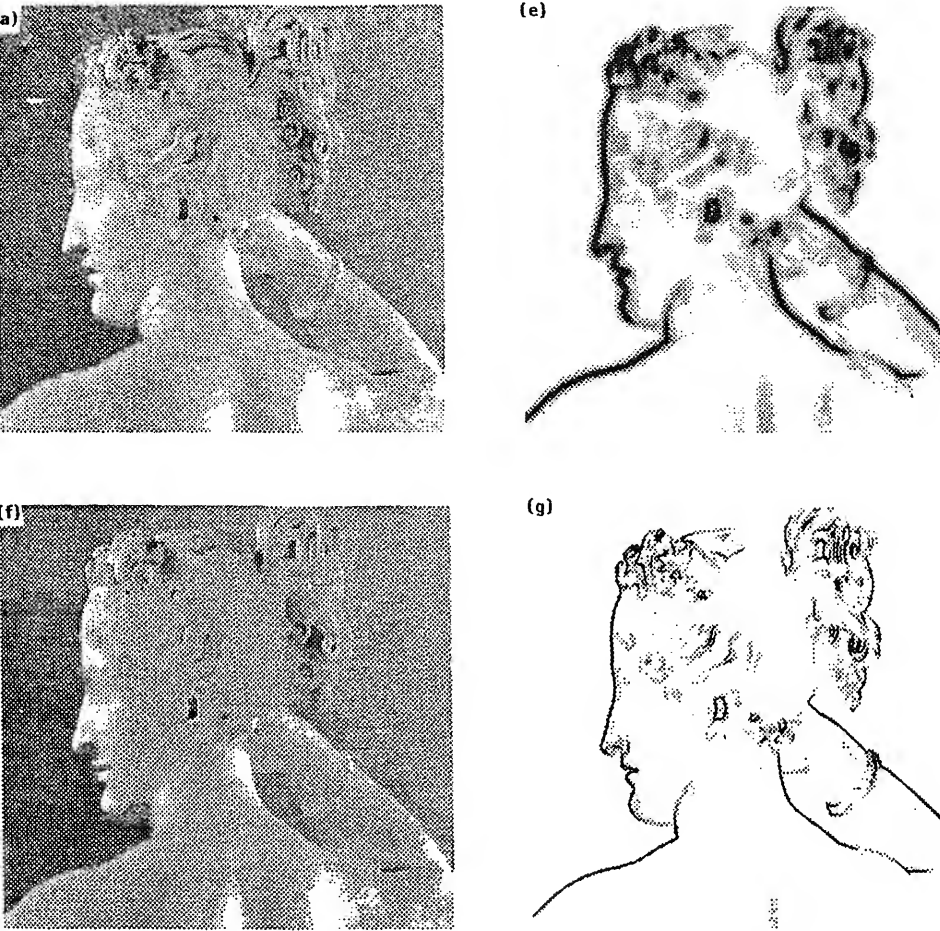
Figure 5. (a) Statue image (Paolina Borghese, Canova circa 1800) 480×512. Prescaling $\Phi(v)$ (**b**), final $\Phi(v)$, (**c**), and final f (**d**), with initial parameters $\beta = 0.04$, $\alpha = 0.01$, $c(0) = 1.0$ and final parameters $\beta = 0.2$, $c = 0.2$. Prescaling $\Phi(v)$ (**e**) final $\Phi(v)$ (**f**), and final f (**g**), with initial parameters $\beta = 0.1$, $\alpha = 0.01$, $c(0) = 1.0$ and final parameters $\beta = 0.5$, $c = 0.2$.

le. The parameters α and β are an order of magnitude smaller in figure 6. This admits ich greater smoothing of the image. Even so, edge localisation is accurate. In both cases displayed images are the following. Figures (b), (c), and (d) are $\Phi(v(0))$, $f(T)$, and $v(T)$) respectively. Figures (e), (f), and (g) reiterate (b), (c), and (d) for the second set parameters in each case. In figure 5 the first set of parameters is given by

$$\beta(0) = 0.1, \quad \alpha = 0.25, \quad c(0) = 2.0, \quad \beta(T) = 1.0, \quad c(T) = 0.2,$$

l the second by

$$\beta(0) = 0.1, \quad \alpha = 0.1, \quad c(0) = 2.0, \quad \beta(T) = 1.0, \quad c(T) = 0.2.$$

figure 6 the first set of parameters is given by

$$\beta(0) = 0.0025, \quad \alpha = 0.05, \quad c(0) = 2.0, \quad \beta(T) = 0.025, \quad c(T) = 0.2,$$
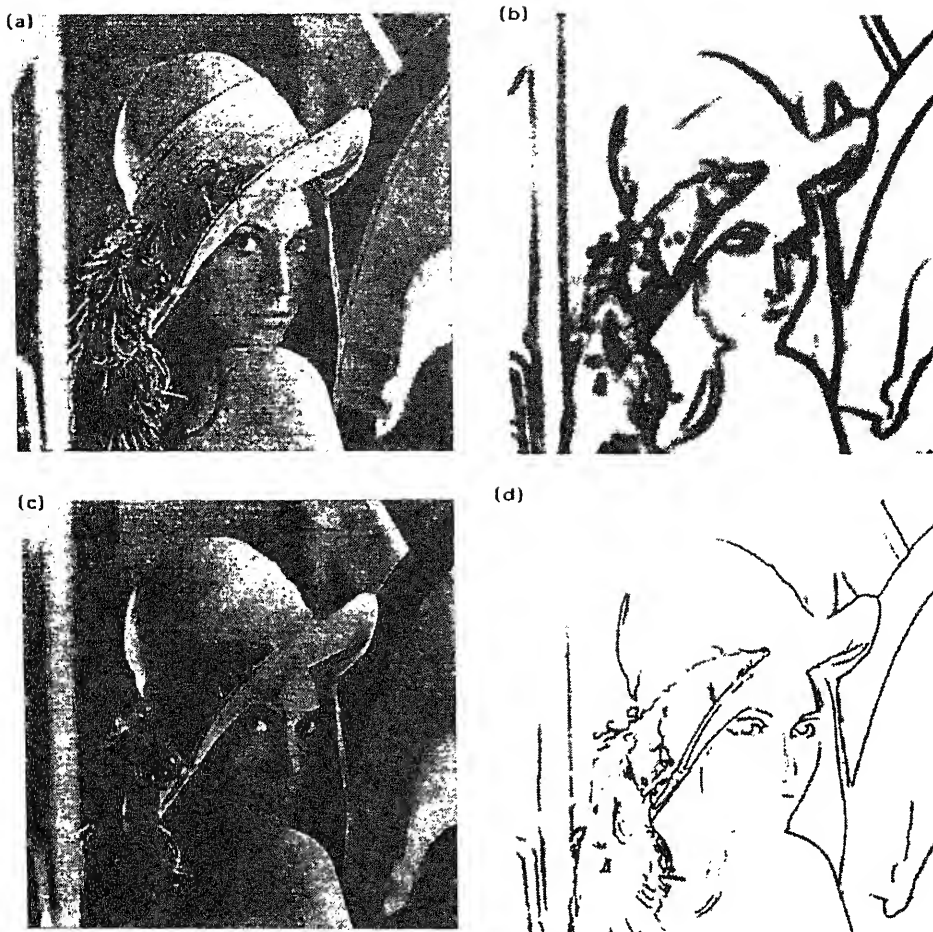
Figure 6. (a–d) (Caption on facing page.)

and the second by

$$\beta(0) = 0.004, \quad \alpha = 0.03, \quad c(0) = 2.0, \quad \beta(T) = 0.04, \quad c(T) = 0.2.$$

We observe that as scale increases the set of edges detected increase monotonically – virtually no scale-dependent distortion is visible.

Figure 6. (a) Lenna image 512×512. Prescaling $\Phi(v)$ (**b**), final f (**c**), and final $\Phi(v)$ (**d**), with initial parameters $\beta = 0.0025$, $\alpha = 0.05$, $c(0) = 2.0$ and final parameters $\beta = 0.025$, $c = 0.2$. Prescaling $\Phi(v)$ (**e**), final f (**f**), and final $\Phi(v)$ (**g**), with initial parameters $\beta = 0.004$, $\alpha = 0.03$, $c(0) = 2.0$ and final parameter $\beta = 0.04$, $c = 0.2$.

References

brosio L 1989 Variational problems on SBV. *Acta Appl. Math.* 17: 1–40

brosio L, Tortorelli V 1990 Approximation of functionals depending on jumps by elliptic unctionals via Γ–convergence. *Commun. Pure Appl. Math.* 43: 999–1036

brosio L, Tortorelli V 1992 On the approximation of functionals depending on jumps by elliptic functionals. *Boll. Un. Mat. Ital.* 7: 105–123

ouch H 1984 *Variational convergence for functions and operators* (London, UK: Pitman)

gholm F 1987 Edge Focusing. *IEEE Trans. Pattern Anal. Machine Intell.* 9: 726–741

ro G L, Snyder W E, Garnier S J, Gault J W 1992 Mean field annealing: A formalism for constructing GNC-like algorithms. *IEEE Trans. Neural Networks* 3: 131–138

ke A, Zisserman A 1987 *Visual reconstruction* (Cambridge, MA: MIT Press)

ny J 1986 A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intell.* 8: 679–698

De Giorgi E, Franzoni T 1979 Su un tipo di convergenze variazionale. *Ren. Sem. Mat. brescia* 3: 63–101

Gamble E B, Poggio T 1987 Visual integration and detection of discontinuities: The key role of intensity edges. A I Memo No. 970, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA

Geiger D, Girosi F 1989 Parallel and deterministic algorithms for MRFs: surface reconstruction and integration, Memo No. 1114

Geiger D, Yuille A 1989 A common framework for image segmentation. Tech. Rep. no. 89–7, Harvard Robotics Laboratory, Harvard University, Cambridge, MA

Geman S, Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* 6: 721–741

Grimson W E L 1981 *From images to surfaces* (Cambridge, MA: MIT Press)

Horn B K T 1986 *Robot vision* (Cambridge, MA: MIT Press)

Harris J, Koch C, Luo J, Wyatt J, 1989 Resistive fuses: Analog hardware for detecting discontinuities in early vision. *Analog VLSI implementations of neural systems* (Norwell, MA: Kluwer) pp 27–55

March R 1988 Computation of stereo disparity using regularization. *Pattern Recognition Lett.* 8: 181–187

March R 1989 A regularization model for stereo vision with controlled continuity. *Pattern Recognition Lett.* 10: 259–263

March R 1992 Visual reconstruction with discontinuities using variational methods. *Image Vision Comput.* 10:

Marr D, Hildreth E 1980 Theory of edge detection. *Proc. R. Soc. London B* 207: 187–217

Marroquin J L 1985 *Probabilistic solution of inverse problems.* Ph D thesis, Dept. of Electrical Engineering and Computer Science, Massachussetts Institute of Technology, Cambridge, MA

Mumford D, Shah J 1985 Boundary detection by minimizing functionals. *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco

Mumford D, Shah J 1989 Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* 42: 577–685

Nordström K N 1990 Biased anisotropic diffusion: a unified regularization and diffusion approach to edge detection. *Image Vision Comput.* 8: 318–327

Perona P, Malik J 1990 Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Machine Intell.* 12: 629–639

Richardson T J 1990 *Scale independent piecewise smooth segmentation of images via variational methods*, Ph D thesis, Dept. of Electrical Engineering and Computer Science, Massachussetts Institute of Technology, Cambridge, MA

Richardson T J 1992 Limit theorems for a variational problem arising in computer vision. *Annali della Scuola Normale* 19: 1–49

Rosenfeld A, Thurston M 1971 Edge and curve detection for visual scene analysis. *IEEE Trans. Comput.* C-20: 562–569

Shah J 1991 Segmentation by non-linear diffusion. *Proc. IEEE Comput. Vision Pattern Recognition 91, Hawai*

Shah J 1992 Segmentation by minimizing functionals: Smoothing properties. *SIAM J. Control Optimization* 30: 99–111

Witkin A 1983 Scale-space filtering. *International Joint Conference on Artificial Intelligence* Karlsruhe, pp 1019–1021

atrix partitioning methods for interior point algorithms

ROMESH SAIGAL

Department of Industrial and Operations Engineering, The University of
Michigan, Ann Arbor, Michigan 48109-2117, USA
e-mail: rsaigal@engin.umich.edu

Abstract. We consider here a linear programming problem whose rows of
the constraint matrix can be partitioned into two parts. Such natural partitions
exist in several special linear programs, including the assignment problem, the
transportation problem, the generalized upper-bounded variable problem, the
block diagonal linear program; and can also be used to induce sparsity pat-
terns in Cholesky factorizations. In this paper, we propose a matrix partitioning
method for interior point algorithms. The proposed method independently gen-
erates Cholesky factorizations of each part, and reduces the complexity to that
of solving generally, a dense linear system involving several rank one updates
of the identity matrix. Here, we propose solving this linear system by an induc-
tive use of the Sherman–Morrison–Woodbury formula. The proposed method
is easily implemented on a vector, parallel machine as well as on a distributed
system. Such partitioning methods have been popular in the context of the
simplex method, where the special structure of the basis matrix is exploited.

Keywords. Linear programming; interior point methods; Cholesky factor-
izations; matrix partitioning methods; vector-parallel machines; distributed
systems.

Introduction

e consider here the linear programming problem

$$\min \quad \mathbf{c}^T \mathbf{x}$$
$$\mathbf{A}\mathbf{x} = \mathbf{b}$$
$$\mathbf{x} \geq \mathbf{0},$$

ere $\mathbf{A}$ is a $m \times n$ matrix of rank m, $\mathbf{b}$ is an m vector and $\mathbf{c}$ is an n vector. We assume
t the matrix $\mathbf{A}$ has the partition

$$\mathbf{A} = \begin{bmatrix} \mathbf{G} \\ \mathbf{H} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \end{bmatrix}, \tag{1}$$

ere $\mathbf{G}$ and $\mathbf{H}$, respectively, are $m_1 \times n$ and $m_2 \times n$ matrices, and $\mathbf{g}$ and $\mathbf{h}$ are, respectively,
and m_2 vectors, with $m_1 + m_2 = m$. We also assume that such a partition is determined

either by the structure of the problem or has been induced by sparsity considerations. Specially structured linear programs that have this *natural* partition include the assignment problem, the transportation problem, the generalized upper bounded variable problem and the block diagonal linear program. Such partitions may also be considered to arise in situations where rows are added, iteratively, to a linear program, and one of the set of rows **G** or **H** may be considered to be the ones added.

Many variants of the simplex method exist which exploit the induced special structure of the basis matrix. We refer the reader to Lasdon (1970) for a background on these variants. Our goal, in this paper, is to consider the application of the recent interior point methods to this partitioned linear program. The study of these methods was started by the seminal work of Karmarkar (1984). Some notable papers that are related to this work are Barnes (1986), Kojima *et al* (1989) and Vanderbei *et al* (1986). The reader is encouraged to browse through these, the recent book by Fang & Puthenpura (1997), and many references therein, for an introduction to these methods. Our starting point in this paper is the implementation of these methods, which requires the solution of a linear system

$$\mathbf{C}\mathbf{z} = \mathbf{d} \tag{2}$$

where $\mathbf{C} = \mathbf{A}\mathbf{D}\mathbf{A}^T$ for some diagonal matrix $\mathbf{D}$; and $\mathbf{d}$, as a function of $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{D}$, is determined by the particular interior point method employed. In this paper, we reduce (2) to solving an $m_2 \times m_2$ system of the form,

$$(\mathbf{I} - \mathbf{E}\mathbf{E}^T)\mathbf{u} = \mathbf{q}, \tag{3}$$

where $\mathbf{E}$ is an $m_2 \times m_1$ matrix. If $\mathbf{E}_j$ is the jth column of the matrix $\mathbf{E}$, it can be readily seen that

$$\mathbf{I} - \mathbf{E}\mathbf{E}^T = \mathbf{I} - \sum_{i=1}^{m_1} \mathbf{E}_j \mathbf{E}_j^T, \tag{4}$$

and we view (4) as m_1 rank one updates to the identity matrix $\mathbf{I}$. We solve the system (4) by an inductive version of the Sherman–Morrison–Woodbury formula. This inductive method requires $O(m_2 m_1^2)$ multiplications, and is advantageous over inverting $\mathbf{I} - \mathbf{E}\mathbf{E}^T$ when $m_2 > m_1$. Another advantage of this method is its ready implementation on a vector and a parallel/distributed computing environment.

In § 2 we present partial Cholesky factorization, the basic idea behind the partitioning technique; and, in § 3 we present a technique to handle several rank one updates. In § 4 we present the variant to handle the transportation and assignment problems, in § 5 the GUB and in § 6 the block diagonal linear program and the multicommodity flow problem. Finally in § 7, we give some preliminary computational results comparing the transportation variant with LOQO, Vanderbei (1992).

2.   Partial Cholesky factorization of $\mathbf{A}\mathbf{D}\mathbf{A}^T$

In this section we generate the general theory we will use to exploit the partition (1) of $\mathbf{A}$. The main result we need for this purpose is the following theorem:

ue $m \times m$ lower triangular matrix L with positive diagonal entries, such that $C = LL^T$.
 use C is dense, $\frac{1}{6}m^3 + \frac{1}{2}m^2 - \frac{3}{2}m$ multiplications are required to compute L.

f. See George & Liu (1981). ∎

he matrix L in the above theorem is called the Cholesky factor of C. Given that sparse, we refer the reader to George & Liu (1981), an excellent reference on the nodology that preserves sparsity of L. We now use this theorem to exploit the partition f the rows of A.

ssuming this partition, it can be easily seen that

$$ADA^T = \begin{bmatrix} GDG^T & GDH^T \\ HDG^T & HDH^T \end{bmatrix}. \tag{5}$$

nen we can prove:

ma 1. There exist lower triangular $m_1 \times m_1$ matrix L_1 and $m_2 \times m_2$ matrix L_3 such $GDG^T = L_1L_1^T$ and $HDH^T = L_3L_3^T$.

f. Since A is full row rank, so are G and H. Since D has positive diagonal entries, G^T and HDH^T are symmetric positive definite matrices, and we have our result from rem 1. ∎

Ve now use the lower triangular matrices L_1 and L_3, guaranteed by lemma 1, for solving vhen A has the partition (1). This is done in the next theorem, which generates a partial lesky factor of ADA^T.

orem 2. Let the $m_1 \times m_1$ matrix L_1 and $m_2 \times m_2$ matrix L_3 be defined as in the na 1. Then, for the $m_2 \times m_1$ matrix L_2 and $m_2 \times m_2$ matrix $\bar{D}$ with

$$\bar{D} = L_3(I - L_3^{-1}L_2L_2^TL_3^{-T})L_3^T, \tag{6}$$

$$L_1L_2^T = GDH^T, \tag{7}$$

$$L = \begin{bmatrix} L_1 & 0 \\ L_2 & I \end{bmatrix}, \quad \hat{D} = \begin{bmatrix} I & 0 \\ 0 & \bar{D} \end{bmatrix},$$

$$ADA^T = L\hat{D}L^T.$$

f. Can be readily verified by a direct multiplication of L, $\hat{D}$ and L^T. ∎

s a result of this theorem, the solution of (2) can be expressed in terms of the matrices L_2, and L_3. This is done in the next result.

ssume that the partitions

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix},$$

form to the partition (5) of C ($= ADA^T$). Here, d_1, z_1 are m_1 vectors and d_2, z_2 are vectors.

orem 3. Let the $m_2 \times m_1$ matrix E be defined so that $L_3E = L_2$. Equation (2) can olved by the following sequence of steps.

Step 1. Find $\mathbf{z}'_1$ *by forward solve*

$$\mathbf{L}_1\mathbf{z}'_1 = \mathbf{d}_1.$$

Step 2. Define

$$\mathbf{z}'_2 = \mathbf{d}_2 - \mathbf{L}_2\mathbf{z}'_1.$$

Step 3. Find $\mathbf{z}''_2$ *by forward solve*

$$\mathbf{L}_3\mathbf{z}''_2 = \mathbf{z}'_2.$$

Step 4. Solve

$$(\mathbf{I} - \mathbf{E}\mathbf{E}^T)\mathbf{z}'''_2 = \mathbf{z}''_2. \tag{8}$$

Step 5. Find $\mathbf{z}_2$ *by backward solve*

$$\mathbf{L}_3^T\mathbf{z}_2 = \mathbf{z}'''_2.$$

Step 6. Find $\mathbf{z}_1$ *by backward solve*

$$\mathbf{L}_1^T\mathbf{z}_1 = \mathbf{z}'_1 - \mathbf{L}_2^T\mathbf{z}_2.$$

Proof. Using the structure of $\mathbf{L}$ and $\hat{\mathbf{D}}$ (of theorem 2) the following equations can be readily derived

$$\mathbf{L}_1\mathbf{z}'_1 = \mathbf{d}_1,$$
$$\mathbf{z}'_2 = \mathbf{d}_2 - \mathbf{L}_2^T\mathbf{z}'_1,$$
$$\bar{\mathbf{D}}\mathbf{z}_2 = \mathbf{z}'_2,$$
$$\mathbf{L}_1^T\mathbf{z}_1 = \mathbf{z}'_1 - \mathbf{L}_2^T\mathbf{z}_2.$$

The theorem now follows from the structure (6) of $\bar{\mathbf{D}}$.

The result of theorem 3 requires the Cholesky factors of $\mathbf{G}\mathbf{D}\mathbf{G}^T$ and $\mathbf{H}\mathbf{D}\mathbf{H}^T$, which are expected to be less dense than the factor of $\mathbf{A}\mathbf{D}\mathbf{A}^T$. The price for this enhanced sparsity is paid at step 4 of the theorem. Here, generally, a dense system of equations has to be solved. In the next section, we will present a procedure that solves this system in $O(m_1^2m_2)$ multiplications. If this system were solved by Cholesky factorization, calculation of $\mathbf{E}\mathbf{E}^T$ would require $m_1m_2^2$ multiplications, and the calculation of the Cholesky factor another $\frac{1}{6}m_2^3 + \frac{1}{2}m_2^2 - \frac{3}{2}m_2$. In the case $m_2 \geq m_1$, solving directly is more expensive.

For a dense matrix $\mathbf{C}$, system (2) can be solved, using Cholesky factors, in $\frac{1}{6}m^3 + \frac{3}{2}m^2 - \frac{5}{2}m$ multiplications. Using the technique of this section with a partitioned matrix $\mathbf{A}$, it requires $\frac{1}{6}(m^3 + 9m^2 + 6m_1^2m_2 - 3m)$ multiplications, and is not competitive.

We now establish a basic property of the matrix $\mathbf{E}\mathbf{E}^T$, encountered in step 4.

Theorem 4. *Let* $\mathbf{E}$ *be defined as in theorem* 3, *step* 4. *The spectral radius of* $\mathbf{E}\mathbf{E}^T$ *is less than* 1.

Proof. Since $\mathbf{A}\mathbf{D}\mathbf{A}^T$ is positive definite, so is $\mathbf{L}\hat{\mathbf{D}}\mathbf{L}^T$ (see theorem 2 for definitions), and thus $\hat{\mathbf{D}}$ is positive definite. Using the structure of $\hat{\mathbf{D}}$, $\bar{\mathbf{D}}$ is positive definite. Thus, from (6), $\mathbf{z}^T(\mathbf{I} - \mathbf{E}\mathbf{E}^T)\mathbf{z} > 0$ for all $\mathbf{z}$. Thus $\mathbf{z}^T\mathbf{E}\mathbf{E}^T\mathbf{z} < \mathbf{z}^T\mathbf{z}$ for all $\mathbf{z}$, and we have our result.

A method for a system with several rank-one updates

his section we develop an inductive version of the Sherman–Morrison–Woodbury
nula for solving the system (8). For each $j = 1, \ldots, m_1$, let $\mathbf{E}_j$ be the jth column of
natrix $\mathbf{E}$, and

$$\mathbf{E}_{m_1+1} = \mathbf{z}_2''.$$

, let $\mathbf{B}_1 = \mathbf{I}$, and for each $k = 1, \ldots, m_1$

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \mathbf{E}_k \mathbf{E}_k^T,$$

note that $\mathbf{B}_{m_1+1} = \mathbf{I} - \mathbf{E}\mathbf{E}^T$.
ow, for each $k = 1, \ldots, m_1 + 1$ and $j = 1, \ldots, m_1 + 1$, define

$$\mathbf{B}_k \mathbf{E}_j^{(k)} = \mathbf{E}_j.$$

s, for each $k = 1, \ldots, m_1$ and $j = k + 1, \ldots, m_1 + 1$,

$$\mathbf{B}_{k+1} \mathbf{E}_j^{(k+1)} = \mathbf{E}_j,$$

$$(\mathbf{B}_k - \mathbf{E}_k \mathbf{E}_k^T) \mathbf{E}_j^{(k+1)} = \mathbf{E}_j,$$

$$(\mathbf{I} - \mathbf{E}_k^{(k)} \mathbf{E}_k^T) \mathbf{E}_j^{(k+1)} = \mathbf{E}_j^{(k)}. \tag{9}$$

Ising the Morrison–Woodbury formula, we can write the solution to (9) as

$$\mathbf{E}_j^{(k+1)} = \mathbf{E}_j^{(k)} + \frac{\langle \mathbf{E}_j, \mathbf{E}_k^{(k)} \rangle}{1 - \langle \mathbf{E}_k, \mathbf{E}_k^{(k)} \rangle} \mathbf{E}_k^{(k)}$$

$$= \mathbf{E}_j^{(k)} + \frac{\langle \mathbf{E}_k, \mathbf{E}_j^{(k)} \rangle}{1 - \langle \mathbf{E}_k, \mathbf{E}_k^{(k)} \rangle} \mathbf{E}_k^{(k)},$$

ere $\langle \cdot, \cdot \rangle$ is the usual inner product of two vectors.
Ve can prove the following result about the above procedure.

eorem 5. *The solution z_2''' of (8) is $\mathbf{E}_{m_1+1}^{(m_1+1)}$.*

of. This is readily seen since by definition,

$$\mathbf{B}_{m_1+1} \mathbf{E}_{m_1+1}^{(m_1+1)} = \mathbf{E}_{m_1+1}$$

stituting the identities $\mathbf{B}_{m_1+1} = \mathbf{I} - \mathbf{E}\mathbf{E}^T$ and $\mathbf{E}_{m_1+1} = \mathbf{z}_2''$, we obtain the theorem.
he above inductive procedure (on k) suggests the following algorithm in a vector and
stributed/parallel environment.

1. 1. Communicate $\mathbf{E}_j$ to processor $j = 1, \ldots, m_1 + 1$.

2. At processor j, set $\mathbf{E}_j^{(1)} = \mathbf{E}_j$.

3. Set $k = 1$.

Step 2. From processor k, communicate $\mathbf{E}_k^{(k)}$, $\mathbf{E}_k$ and $\langle \mathbf{E}_k, \mathbf{E}_k^{(k)} \rangle$ to each processor $j = k + 1, \ldots, m_1 + 1$.

Step 3. At processor j, $j = k + 1, \ldots, m_1 + 1$ compute

$$\mathbf{E}_j^{(k+1)} = \mathbf{E}_j^{(k)} + \frac{\langle \mathbf{E}_k, \mathbf{E}_j^{(k)} \rangle}{1 - \langle \mathbf{E}_k, \mathbf{E}_k^{(k)} \rangle} \mathbf{E}_k^{(k)}.$$

Step 4. Set $k = k + 1$. If $k \leq m_1$, then go to step 2, otherwise declare $\mathbf{E}_{k+1}^{(k+1)}$ as the solution z_2''' of (8).

A careful count of the number of multiplications needed are summarized in the following theorem.

Theorem 6. *The number of multiplications required by the algorithm to solve* (8) *is* $m_1^2 m_2 + 2m_1 m_2$. *In the vector and parallel environment, the number of vector operations required is* $3m_1$.

Proof. Can be readily verified by a careful counting.

There is a considerable advantage in keeping $m_1 < m_2$ while solving (8), an $m_2 \times m_2$ system, by the above strategy. Otherwise, it would require $m_1 m_2^2$ multiplications to obtain $\mathbf{EE}^T$, $\frac{1}{6}(m_2^3 + 3m_2^2 - 9m_2)$ to obtain the Cholesky factor of $\mathbf{I} - \mathbf{EE}^T$ and $m_2(m_2 + 1)$ to solve the system using the Cholesky factor. The method of this section is less advantageous if $m_2 < m_1$; and to use it effectively m_1 and m_2 should be re-defined.

4. Transportation and assignment problem

The transportation problem is the following: given m supply depots, with s_i (> 0) as the units of supply of some good at depot i for each $i = 1, \ldots, m$; n demand centres with d_j (> 0) as the units of demand of the good at the centre j for each $j = 1, \ldots, n$; and $c_{i,j}$ (> 0) the cost of shipping one unit of good between depot i and centre j for each $i = 1, \ldots, m$, $j = 1, \ldots, n$; find the least cost quantity of good shipped between each depot and centre. We assume here that the transportation problem is balanced, or $\sum_{i=1}^m s_i = \sum_{j=1}^n d_j$; i.e., there are just enough goods available at the supply depots to meet this demand. In this case it is well known that the transportation problem has an optimal shipping schedule, and it can be found by solving the following *Transportation Linear Program*,

$$\min \sum_{i=1}^m \sum_{j=1}^n c_{i,j} x_{i,j}$$

$$\sum_{i=1}^{m} x_{i,j} = d_j, \quad j = 1, \ldots, n,$$

$$x_{i,j} \geq 0, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n,$$

ere $x_{i,j}$ is the number of units of goods shipped from depot i to centre j, for each i
j.

The assignment problem is the following: given there are n individuals that can do
of the n tasks; and, assigning the individual i to the task j costs $c_{i,j} (> 0)$ for each
$1, \ldots, n$, $j = 1, \ldots, n$, find the least cost assignment of individuals to tasks. It is also
uired that each individual perform exactly one task, and that each task be performed
exactly one individual. This problem can be cast as a balanced transportation problem,
h each individual associated with a supply depot with exactly one unit of supply, and
h task with a demand centre with exactly one unit of demand. The only difference is
t the variables $x_{i,j}$ are required to take on values 0 or 1. As is well known, because of
total unimodularity property of the constraint matrix, setting $s_i = 1$ and $d_j = 1$ for
h $i = 1, \ldots, n$ and $j = 1, \ldots, n$ and solving the transportation linear program suffices
ind the optimal assignment.

The dual of this linear program is the following:

$$\max \sum_{i=1}^{m} s_i u_i + \sum_{j=1}^{n} d_j v_j$$

$$u_i + v_j + s_{i,j} = c_{i,j}, \quad \text{for all } i = 1, \ldots m, \quad j = 1, \ldots, n,$$

$$s_{i,j} \geq 0, \quad \text{for all } i = 1, \ldots m, \quad j = 1, \ldots, n,$$

ere u_i and v_j are the dual variables and $s_{i,j}$ are the dual slacks for each i and j.

The constraints of the (primal) linear program fall naturally into two sets, the supply
istraints and the demand constraints. It is this partition that we will exploit in the
orithm. As is well known, the rank of the constraint matrix is $m + n - 1$, which is one
s than the number of constraints. Thus one constraint must be discarded to ensure that
constraint matrix has full row rank. We will discard the supply constraint associated
th the depot m, and use the following constraint matrix:

$$\mathbf{A} = \begin{bmatrix} \mathbf{e}^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}^T & \cdots & \mathbf{0} & \mathbf{0} \\ & & \ddots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e}^T & \mathbf{0} \\ \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} & \mathbf{I} \end{bmatrix},$$

ere $\mathbf{e}^T = (1, 1, \ldots, 1)$, an n vector, and $\mathbf{I}$ is the $n \times n$ identity matrix.
Thus we define,

$$\mathbf{G} = \begin{bmatrix} \mathbf{e}^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}^T & \cdots & \mathbf{0} & \mathbf{0} \\ & & \ddots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e}^T & \mathbf{0} \end{bmatrix}, \qquad \mathbf{H} = (\mathbf{I}, \mathbf{I}, \ldots, \mathbf{I}),$$

where $\mathbf{G}$ is an $(m-1) \times mn$ and $\mathbf{H}$ is an $n \times mn$ matrix. The diagonal matrix $\mathbf{D}$ has mn entries along the diagonal, and has the partition

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}^{(1)} & 0 & \cdots & 0 \\ 0 & \mathbf{D}^{(2)} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \mathbf{D}^{(m)} \end{bmatrix},$$

where $\mathbf{D}^{(i)}$ is an $n \times n$ diagonal matrix, and its jth diagonal entry $D_{jj}^{(i)}$ is the ijth diagonal entry of $\mathbf{D}$, which corresponds to the entry of $\mathbf{D}$ associated with the variable $x_{i,j}$. The exact form of this entry depends on the particular interior point algorithm implemented.

It can now be readily confirmed that

$$\mathbf{GDG}^T = \begin{bmatrix} \mathbf{e}^T \mathbf{D}^{(1)} \mathbf{e} & & & \\ & \mathbf{e}^T \mathbf{D}^{(2)} \mathbf{e} & & \\ & & \ddots & \\ & & & \mathbf{e}^T \mathbf{D}^{(m-1)} \mathbf{e} \end{bmatrix},$$

an $(m-1) \times (m-1)$ diagonal matrix; and

$$\mathbf{HDH}^T = \sum_{i=1}^{n} \mathbf{D}^{(i)}$$

an $n \times n$ diagonal matrix. Thus $\mathbf{L}_1$ and $\mathbf{L}_3$ are diagonal matrices, with the jth diagonal entry of $\mathbf{L}_1$ and $\mathbf{L}_3$ being $(\sum_{k=1}^{n} D_{kk}^{(j)})^{\frac{1}{2}}$ and $(\sum_{i=1}^{m} D_{jj}^{(i)})^{\frac{1}{2}}$, respectively. Also

$$\mathbf{GDH}^T = \begin{bmatrix} \mathbf{e}^T \mathbf{D}^{(1)} \\ \mathbf{e}^T \mathbf{D}^{(2)} \\ \vdots \\ \mathbf{e}^T \mathbf{D}^{(m-1)} \end{bmatrix},$$

and

$$\mathbf{L}_2 = (\mathbf{D}^{(1)}\mathbf{e}, \mathbf{D}^{(2)}\mathbf{e}, \ldots, \mathbf{D}^{(m-1)}\mathbf{e})\mathbf{L}_1^{-1}.$$

Thus

$$\begin{aligned} \mathbf{E} &= \mathbf{L}_3^{-1}\mathbf{L}_2 \\ &= \mathbf{L}_3^{-1}(\mathbf{D}^{(1)}\mathbf{e}, \mathbf{D}^{(2)}\mathbf{e}, \ldots, \mathbf{D}^{(m-1)}\mathbf{e})\mathbf{L}_1^{-1}, \end{aligned}$$

which is an $n \times (m-1)$ positive matrix.

Theorem 7. *Each iteration of the interior point method requires $nm^2 + 6nm + 2(m-n+1)$ multiplications for the partitioned assignment problem.*

Proof. Using the structure of $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$; and solving (8) by the method of § 3, and carefully counting the multiplications at each step, we get the theorem.

Generalized upper bounding problem

generalized upper bounding linear program is the following:

$$
\begin{array}{llllll}
\text{minimize} & \mathbf{c}_0\mathbf{x}_0 & + & \mathbf{c}_1^T\mathbf{x}_1 & + & \mathbf{c}_2^T\mathbf{x}_2 & + & \cdots & + & \mathbf{c}_p^T\mathbf{x}_p \\
& \mathbf{A}_0\mathbf{x}_0 & + & \mathbf{A}_1\mathbf{x}_1 & + & \mathbf{A}_2\mathbf{x}_2 & + & \cdots & + & \mathbf{A}_p\mathbf{x}_p & = \mathbf{b} \\
& & & \mathbf{e}_1^T\mathbf{x}_2 & & & & & & & = 1 \\
& & & & & \mathbf{e}_2^T\mathbf{x}_2 & & & & & = 1 \\
& & & & & & & \ddots & & & \vdots \\
& & & & & & & & & \mathbf{e}_p^T\mathbf{x}_p & = 1 \\
& \mathbf{x}_0 \geq \mathbf{0} & & \mathbf{x}_1 \geq \mathbf{0} & & \mathbf{x}_2 \geq \mathbf{0} & & \cdots & & \mathbf{x}_p \geq \mathbf{0}
\end{array}
$$

ere, for each $j = 0, \ldots, p$, $\mathbf{A}_j$ is an $m \times n_j$ matrix, $\mathbf{c}_j$ and $\mathbf{x}_j$ are n_j vectors; and for $= 1, \ldots, p$, $\mathbf{e}_j$ a is an n_j vector of all ones. Transportation and assignment problems e this structure as well, but in important applications, $m << p$ (i.e., m is much less p). We make the usual assumption that the constraint matrix has the full row rank $+ p$.

As is evident, the constraints of this problem have the *natural* partition (1) with

$$
\mathbf{G} = (\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_p)
$$

$$
\mathbf{H} = \begin{bmatrix}
0 & \mathbf{e}_1^T & 0 & \cdots & 0 \\
0 & 0 & \mathbf{e}_2^T & \cdots & 0 \\
& & & \ddots & \\
0 & 0 & 0 & \cdots & \mathbf{e}_p^T
\end{bmatrix},
$$

nd the diagonal matrix $\mathbf{D}$ has the partition (corresponding to the partiton of $\mathbf{A}$)

$$
\mathbf{D} = \begin{bmatrix}
\mathbf{D}^{(0)} & & & \\
& \mathbf{D}^{(1)} & & \\
& & \ddots & \\
& & & \mathbf{D}^{(p)}
\end{bmatrix}, \tag{10}
$$

vhere $\mathbf{D}^{(j)}$ is an $n_j \times n_j$ diagonal matrix corresponding to columns of $\mathbf{x}_j$ in $\mathbf{A}$, for each $= 1, \ldots, p$.

Here

$$
\mathbf{GDG}^T = \sum_{j=0}^{p} \mathbf{A}_j \mathbf{D}^{(j)} \mathbf{A}_j^T,
$$

$$
\mathbf{HDH}^T = \begin{bmatrix}
\mathbf{e}_1^T\mathbf{D}^{(1)}\mathbf{e}_1 & & & \\
& \mathbf{e}_2^T\mathbf{D}^{(2)}\mathbf{e}_2 & & \\
& & \ddots & \\
& & & \mathbf{e}_p^T\mathbf{D}^{(p)}\mathbf{e}_p
\end{bmatrix},
$$

with $\mathbf{HDH}^T$ a diagonal matrix. Thus $\mathbf{L}_3$ is a diagonal matrix, and $\mathbf{L}_1$, the Cholesky factor of the $m \times m$ matrix

$$\sum_{j=0}^{p} \mathbf{A}_j \mathbf{D}^{(j)} \mathbf{A}_j^T = \mathbf{L}_1 \mathbf{L}_1^T.$$

In applications, even when $\mathbf{A}_j$ are sparse, we would expect their sum to be considerably more dense, and thus we expect $\mathbf{L}_1$ to be relatively dense. Also

$$\mathbf{GDH}^T = (\mathbf{A}_1 \mathbf{D}^{(1)} \mathbf{e}_1, \mathbf{A}_2 \mathbf{D}^{(2)} \mathbf{e}_2, \ldots, \mathbf{A}_p \mathbf{D}^{(p)} \mathbf{e}_p).$$

Thus if $\mathbf{L}_2 = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)$,

$$\mathbf{L}_1 \mathbf{l}_j = \mathbf{A}_j \mathbf{D}^{(j)} \mathbf{e}_j,$$

and we can expect the $p \times m$ matrix $\mathbf{L}_2$ to be dense. In addition, $\mathbf{E} = \mathbf{L}_3^{-1} \mathbf{L}_2$.

Theorem 8. *Each step of an interior point method requires $\frac{3}{2}pm^2 + \frac{1}{6}m^3 + \frac{11}{2}pm + \frac{3}{2}m^2 - \frac{1}{2}m + 2p$ when applied to the partitioned GUB problem.*

Proof. Can be obtained by a careful calculation of the work at each step of the algorithm. The multiplications required to obtain the Cholesky factor $\mathbf{L}_1$ are included in the above formula.

6. Block diagonal linear program

The block diagonal linear program is the following:

$$
\begin{array}{llllllll}
\text{minimize} & \mathbf{c}_0^T \mathbf{x}_0 & + & \mathbf{c}_1^T \mathbf{x}_1 & + & \mathbf{c}_2^T \mathbf{x}_2 & + \cdots + & \mathbf{c}_p^T \mathbf{x}_p \\
& \mathbf{A}_0 \mathbf{x}_0 & + & \mathbf{A}_1 \mathbf{x}_1 & + & \mathbf{A}_2 \mathbf{x}_2 & + \cdots + & \mathbf{A}_p \mathbf{x}_p & = \mathbf{b}_0 \\
& & & \bar{\mathbf{A}}_1 \mathbf{x}_2 & & & & & = \mathbf{b}_1 \\
& & & & & \bar{\mathbf{A}}_2 \mathbf{x}_2 & & & = \mathbf{b}_2 \\
& & & & & & \ddots & & \vdots \\
& & & & & & & \bar{\mathbf{A}}_p \mathbf{x}_p & = \mathbf{b}_p \\
& \mathbf{x}_0 \geq 0 & & \mathbf{x}_1 \geq 0 & & \mathbf{x}_2 \geq 0 & \cdots & \mathbf{x}_p \geq 0
\end{array}
$$

where, for each $j = 0, \ldots, p$, $\mathbf{A}_j$ is an $m_0 \times n_j$ matrix, $\mathbf{c}_j$ and $\mathbf{x}_j$ are n_j vectors; and, for each $j = 1, \ldots, p$, $\bar{\mathbf{A}}_j$ is an $m_j \times n_j$ matrix. Let $m = m_0$ and $M = m_1 + \cdots + m_p$. The constraints of this problem have the *natural* partition (1) with

$$\mathbf{G} = (\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_p)$$

$$
\mathbf{H} = \begin{bmatrix}
0 & \bar{\mathbf{A}}_1 & 0 & \cdots & 0 \\
0 & 0 & \bar{\mathbf{A}}_2 & \cdots & 0 \\
& & & \ddots & \\
0 & 0 & 0 & \cdots & \bar{\mathbf{A}}_p
\end{bmatrix},
$$

and the diagonal matrix $\mathbf{D}$ has the partition (10). It can be verified that

$$\mathbf{GDG}^T = \sum_{j=0}^{p} \mathbf{A}_j \mathbf{D}^{(j)} \mathbf{A}_j^T;$$

$$\mathbf{HDH}^T = \begin{bmatrix} \bar{\mathbf{A}}_1 \mathbf{D}^{(1)} \bar{\mathbf{A}}_1^T & & & \\ & \bar{\mathbf{A}}_2 \mathbf{D}^{(2)} \bar{\mathbf{A}}_2^T & & \\ & & \ddots & \\ & & & \bar{\mathbf{A}}_p \mathbf{D}^{(p)} \bar{\mathbf{A}}_p^T \end{bmatrix},$$

block diagonal matrix. As in the GUB case, the factor $\mathbf{L}_1$ of $\mathbf{GDG}^T$ will be relatively
e. $\mathbf{L}_3$, the Cholesky factor of $\mathbf{HDH}^T$ is block diagonal, and each diagonal factor can
omputed independently. Thus, we can assume that

$$\mathbf{L}_3 = \begin{bmatrix} \mathbf{L}_{31} & & & \\ & \mathbf{L}_{32} & & \\ & & \ddots & \\ & & & \mathbf{L}_{3p} \end{bmatrix}.$$

hen $\bar{\mathbf{A}}_j$ is sparse, we can preserve the sparsity of $\mathbf{L}_{3j}$ (the Cholesky factor of $\bar{\mathbf{A}}_j \mathbf{D}^{(j)} \bar{\mathbf{A}}_j^T$)
e techniques of sparse Cholesky factorization, George & Liu (1981). Also

$$\mathbf{GDH}^T = (\mathbf{A}_1 \mathbf{D}^{(1)} \bar{\mathbf{A}}_1^T, \dots, \mathbf{A}_p \mathbf{D}^{(p)} \bar{\mathbf{A}}_p^T)$$

$m \times M$ matrix. If $\mathbf{L}_2^T = (\mathbf{L}_{21}^T, \dots, \mathbf{L}_{2p}^T)$,

$$\mathbf{L}_1 \mathbf{L}_{2j}^T = \mathbf{A}_j \mathbf{D}^{(j)} \bar{\mathbf{A}}_j^T.$$

expect $\mathbf{L}_{2j}$ to be, generally, dense.
et $\mathbf{E} = (\mathbf{E}^{(1)}, \mathbf{E}^{(2)}, \dots, \mathbf{E}^{(p)})^T$ which conforms to the partition of $\mathbf{L}_3$, Then, for each
$1, 2, \dots, p$

$$\mathbf{L}_{3j} \mathbf{E}^{(j)} = \mathbf{L}_{2j}$$

we can readily establish the following theorem:

orem 9. *For a dense problem, it takes $\frac{1}{6} \sum_{j=0}^{p} (m_j^3 + 3m_j^2 - 9m_j) + \frac{1}{2} Mm(m + \frac{m}{2} \sum_{j=1}^{p} m_j(m_j + 1)$ multiplications to generate $\mathbf{L}_1$, $\mathbf{L}_2$, $\mathbf{L}_3$ and $\mathbf{E}$; and, takes $^2 + 4Mm + m^2 + m + \sum_{j=1}^{p} m_j^2 + M$ multiplications to solve the systems of the- (4). If $m_j = \bar{m}$ for all $j = 1, \dots, p$, the above forms reduce to $\frac{1}{6}(m^3 + 3m^2 - + pm̄^3 + 3p\bar{m}^2 - 9p\bar{m}) + \frac{1}{2} p\bar{m}m(m + 1) + \frac{1}{2} pm\bar{m}(\bar{m} + 1)$ and $pm^2\bar{m} + 4pm\bar{m} + + m + p\bar{m}^2 + p\bar{m}$ respectively. Thus, in this case, the computations grow linearly*

of. This can be readily proved by a careful counting of the required multiplications.

arc incidence matrix of the directed network through which these commodities flow. Thus,

$$GDG^T = \sum_{j=0}^{p} D^{(j)}$$

is diagonal, and the sparsity pattern of $\bar{A}_i D^{(i)} \bar{A}_i^T$, for each $i = 1, \ldots, p$, is the same. Also,

$$A_i D^{(i)} \bar{A}_i^T = D^{(i)} R^T.$$

Thus

$$L_{2i}^T = \left(\sum_{j=1}^{p} D^{(j)} \right)^{-\frac{1}{2}} D^{(i)} R^T,$$

$$L_{3i} E^{(i)} = R D^{(i)} \left(\sum_{j=1}^{p} D^{(j)} \right)^{-\frac{1}{2}}. \tag{11}$$

L_{3i} is the Cholesky factor of $RD^{(i)} R^T$, and, for each i, has the same sparsity pattern as that of the factor of RR^T. It is readily confirmed that, in the notation of George & Liu (1981), the graph associated with RR^T is the unordered network on which the commodities flow. This facilitates considerably the use of their techniques for generating sparse Cholesky factors L_{3i}. Since each column of R has only two nonzero entries, $E^{(i)}$ will be sparse; and will have the same sparsity pattern for each i.

For a network with m nodes, n arcs and p commodities, L_1 is an $n \times n$ diagonal matrix, for each $i = 1, \ldots, p$, L_{3i} is an $m \times m$ matrix, L_{2i} is an $m \times n$ and E is $pm \times n$. Also, assume that it takes δ multiplications to get a sparse Cholesky factor of RR^T; and note that $\delta \leq \frac{1}{6}(m^3 + 3m^2 - 9m)$; and that the number of non-zero elements in this factor are η. Then, the following can be shown.

Theorem 10. *Given L_1, L_2, L_3 and E, it takes $pmn^2 + 4pn + 2n + p\eta$ multiplications to solve all the systems of theorem 3.*

7. Computational experience

In this section, we present some computational experience of solving assignment problems by the procedure suggested here and by the state-of-the-art code LOQO (Vanderbei 1992).

LOQO is a state-of-the-art interior point code based on the primal-dual homotopy method, and implements a predictor–corrector strategy for tracing the path of centres. The per iteration times of this code are compared with the per iteration times of a specialized transportation code, implementing the dual affine scaling strategy. The LOQO per iteration times may be a little larger because of the step size selection in the predictor-corrector strategy. The per-iteration times and their ratio are given in table 1. We point out

lem	Size	Assignment			LOQO			Ratio Loqo/Asgn
		Iter	Time*	Iter time	Iter	Time	Iter time	
	200×200	14	20.78	1.48	15	75.72	5.048	3.41
	200×200	17	25.09	1.476	17	85.30	4.911	3.33
	200×200	21	31.02	1.477	20	96.61	4.830	3.27
	300×300	14	65.72	4.69	18	301.07	16.72	3.565
	300×300	19	88.80	4.67	19	314.16	16.53	3.540
	300×300	21	98.06	4.67	21	418.5	16.86	3.61

is time is the total time for all the iterations, without the input/output time. All times are in seconds.

for these special problems, the specialized version was about 3 to 4 times faster than
general purpose code, LOQO.

he three problems of size 200 × 200; and, of size 300 × 300 are generated randomly,
present increasing difficulty to interior point methods. On the first problem, these
hods converge to a solution in the interior of a face, with very few variables at value 1.
the second problem, these methods converge to a solution with more than 75% of the
ables at value 1, while on the third problem they converge to a vertex, with all variables
alue 1. For the first problem, LOQO found a solution to the accuracy of eight significant
res, but for the second and third, it was unable to find this accurate a solution. For these
olems it found a solution to 7 digits of accuracy.

erences

nes E R 1986 A variation of Karmarkar's algorithm for solving linear programming problems. *Math. Program.* 36: 174–182

g S C, Puthenpura S 1997 *Linear optimization and extensions* (Englewood Cliffs, NJ: Prentice Hall) (in press)

rge A, Liu J W 1981 *Computer solution of large positive definite systems* (Englewood Cliffs, NJ: Prentice Hall)

markar N 1984 A new polynomial time algorithm for linear programming *Combinatorica* 4: 73–395

ima M, Mizuno S, Yoshise A 1989 A primal-dual interior point method for linear programming. In *Progress in mathematical programming: Interior point methods* (ed.) N Megiddo (New York: Springer-Verlag) pp 29–48

don L S 1970 *Optimization theory for large systems* (London: Macmillan)

derbei R J 1992 LOQO user's manual. Program in statistics and operations research. Princeton University, Princeton, NJ

derbei R J, Meketon M S, Freedman B A 1986 A modification of Karmarkar linear programming algorithm. *Algorithmatica* 1: 395–407

parental product algorithm
coded waveform design in radar

P S MOHARIR [1], V M MARU [1] and R SINGH [2]

[1] National Geophysical Research Institute, Uppal Road, Hyderabad
500 007, India
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh,
PA 15213-3890, USA
e-mail: postmast@csngri.ren.nic.in

Abstract. The problem of obtaining long sequences with finite alphabet and peaky aperiodic auto-correlation is important in the context of radar, sonar and system identification and is called the coded waveform design problem, or simply the signal design problem in this limited context. It is good to remember that there are other signal design problems in coding theory and digital communication. It is viewed as a problem of optimization. An algorithm based on two operational ideas is developed. From the earlier experience of using the eugenic algorithm for the problem of waveform design, it was realised that rather than random but multiple mutations, all the first-order mutations should be examined to pick up the best one. This is called Hamming scan, which has the advantage of being locally complete, rather than random. The conventional genetic algorithm for non-local optimization leaves out the anabolic role of chemistry of allowing quick growth of complexity. Here, the Hamming scan is made to operate on the Kronecker or Chinese product of two sequences with best-known discrimination values, so that one can go to large lengths and yet get good results in affordable time. The details of the ternary pulse compression sequences obtained are given. They suggest the superiority of the ternary sequences.

Keywords. Coded waveform design; global optimization; bi-parental products; Hamming scan.

Introduction

term signal design has different meanings in coding theory, spread-spectrum comunication, and radar. Here it is short for a coded waveform design problem in radar.

This involves obtaining sequences of finite lengths with small discrete alphabet, which are good approximations to white noise (Golay 1977). In particular, the alphabet $(-1, 0, 1)$ is considered. The sequences using them are called ternary sequences. Much of the earlier work deals with binary sequences with $(-1, 1)$ as the alphabet. The goodness of the degree of approximation is measured here in terms of the discrimination. An alternative measure is the merit factor (Golay 1977).

To set the notation and to concretize the ideas, let

$$\mathbf{s} = [s_0, s_1, \ldots, s_{n-2}, s_{n-1}] \tag{1}$$

be the sequence of length n, where the elements s_i are from any of the two alphabets referred to above. Then

$$\rho(k) = \sum_{i=0}^{n-1-k} s_i s_{i+k}, \quad k = 0, 1, \ldots, n-1, \tag{2}$$

is called the aperiodic auto-correlation of the sequence $\mathbf{s}$. The quantity,

$$D = \rho(0) \Big/ \left[\max_{k \neq 0} |\rho(k)| \right], \tag{3}$$

is called the discrimination.

The problem of obtaining sequences with high discrimination is a very difficult one. It is so because whereas for the binary sequences with good periodic auto-correlation, analytical design procedures or constructions (Baumert 1971; Golay 1983; Hoholdt *et al* 1985; Reddy & Rao 1986; Jensen *et al* 1991) based on powerful number-theoretical results are available, for sequences with good aperiodic autocorrelations it has still to be basically a search. Further, the search cannot be exhaustive, because the number of sequences to be searched for grows combinatorially, as the length increases. Here, the emphasis is on obtaining as good sequences as possible with certain reasonably efficient procedures without claiming that even higher discrimination values are not possible. Golay's notion of sieves (Golay 1977) which restricts the search to a subset of sequences which can either be designed or searched efficiently and has properties which are desirable for enhancing the peakiness is a similar one. The sieves promise some good sequences without claiming that all good or even the best sequences are necessarily retained.

There is extensive work on binary sequences (Barker 1953; Turyn 1963, 1968; Boehmer 1967; Baumert 1971; Golay 1972, 1977, 1982, 1983; Moharir 1975; Beenker *et al* 1985; Hoholdt *et al* 1985; Kerdock *et al* 1986; Reddy & Rao 1986; Bernasconi 1987, 1988; Hoholdt & Justesen 1988; Golay & Harris 1990; Newmann & Byrnes 1990; Jensen *et al* 1991; De Groot *et al* 1992), which however we do not attempt to review here. Beyond the Barker sequences (Barker 1953) of length 11 and 13, having the discrimination values of 11 and 13, the four best binary sequences having discrimination values of 14, 15, 16 and 17 (Kerdock *et al* 1986) have been obtained at the smallest lengths of 28, 51, 69 and 88. Complete enumeration is possible only up to small lengths (Kerdock *et al* 1986) and would not be regarded as too satisfactory, the ultimate aim being construction rather than search. Partial searches using the notion of sieves (Golay 1977) were extended to higher lengths. Starting with binary sequences with ideal periodic autocorrelation and searching

ve been obtained. To cite some examples (Reddy & Rao 1986), discriminations of 22.82, .85, 36.86, 41.45, 41.77 and 42.46 have been obtained at the lengths of 251, 349, 811, 9, 919 and 1019, respectively.

Optimization techniques have been used for the signal design problem (Bernasconi 87, 1988; De Groot *et al* 1992), but the design criterion has been different, viz., merit :tor.

Ternary aperiodic sequences have been proposed and listed earlier with merit factor as e criterion of goodness (Moharir 1974, 1976; Moharir *et al* 1985). They were obtained ing sieves such as terminal admissibility (Moharir 1976), skew-symmetry (Golay 1977), minal admissibility for skew-symmetric sequences (Moharir *et al* 1985) etc. This earlier ork establishing the superiority of ternary pulse compression sequences over binary quences, if merit factor is accepted as a valid desideratum, is extended further in this per with discrimination as a chosen criterion.

The algorithm and associated concepts

1 *Hamming scan*

improved version of the genetic algorithm (Holland 1992; Michalewicz 1992) called genic algorithm was used (Singh *et al* 1996) to see whether better and longer ternary quences could be obtained. Success was met with along both these directions. Yet it was t possible to go to very large lengths, as the search time requirement increased very fast. was regarded that one should devise an optimization algorithm which is more efficient, en though possibly more suboptimal. The Hamming scan is one such algorithm. Genetic gorithms use random but possibly multiple mutations. Mutation is a term metaphorically ed for a change in an element in the sequence. Thus, a single mutation is at a Hamming stance of one from the original sequence. The Hamming scan looks at all the Hamming 1 neighbours and picks up the one with the largest discrimination. If it is better than e original sequence with the chosen definition of goodness, the algorithm is recursively ntinued therefrom, as long as improvement is possible. Thus, an entirely probabilistic echanism of mutation is replaced by a locally complete search. The Hamming scan is pedited, and hence, made applicable at large lengths, by not calculating the aperiodic tocorrelation of a Hamming neighbour *ab initio*, recognizing the fact that as only one ement is different, only its different contributions need to be taken into account. Let the ement s_j be changed to c_j. As a result, let $\rho(k)$ change to $\rho'(k)$. Then it can be shown that

$$\rho'(0) = \rho(0) + (c_j^2 - s_j^2), \tag{4}$$

d

$$\rho'(k) = \rho(k) + (c_j - s_j)s_{j+k} + s_{j-k}(c_j - s_j), \quad k = 1, 2, \ldots, n - 1;$$
$$j = 0, 1, \ldots, n - 1. \tag{5}$$

In (5), there are two correction terms. They have to be implemented with care. One ay is to assume that s_p is equal to zero if p is outside $(0, 1, \ldots, n - 1)$. Alternatively, e correction term $s_{j-k}(c_j - s_j)$ is included only for $k = 1, 2, \ldots, n - j - 1$ and the

correction term $s_{j-k}(c_j - s_j)$ is included only for $k = j+1, j+2, \ldots, n-1$. This idea is certainly trivial, but it has led to significantly increased efficiency, and hence, to search at longer lengths than would otherwise have been possible.

For the ternary sequences the actual implementation can be somewhat different. The elements in the sequence are $-1, 0$ and 1. Each one of them can mutate in two possible ways. The mutations $-1 \to 0, 0 \to -1$ and $1 \to -1$ are considered to give the lower strand of Hamming neighbours and the other mutations $-1 \to 1, 0 \to 1$ and $1 \to 0$ are regarded as giving the upper strand of Hamming neighbours. The best neighbours along these two strands were found separately. The idea is that if one strand gives improvement, the other strand may not even be considered in order to save time. As the results are certainly path-dependent and optimality is not guaranteed, efficiency is a valid determinant.

The Hamming scan yielded some better ternary sequences in reasonable time than were obtainable with the eugenic algorithm. However, the Hamming scan also became unaffordable at larger lengths.

2.2 Simon's principle through Kronecker and Chinese products

That is when Simon's principle (Koestler 1969; Simon 1981) suggested newer possibilities. It states that bigger systems evolve faster, when developed through the metastable intermediate subsystems, than if they are constructed *ab initio* from the smallest components. In the present context, it took the form of using two sequences of the best discrimination values available and obtaining a sequence of much larger length from them, such that it already had better discrimination than would result from random choice. The actual mechanism is provided by bi-parental products (Moharir 1992) of two or more sequences. In particular, two products, viz., Kronecker product (Brewere 1978; Moharir 1992) and the Chinese product (Moharir 1977, 1992) are chosen. These products are said to be bi-parental because each element in the product depends exactly on one element each from the two component sequences.

The Kronecker product of two sequences

$$\mathbf{s_1} = [s_{01}, s_{11}, \ldots, s_{(p-1)1}], \qquad \mathbf{s_2} = [s_{02}, s_{12}, \ldots, s_{(q-1)2}], \tag{6}$$

of lengths p and q respectively, is a sequence $\mathbf{s}$ of length pq, defined as

$$\mathbf{s_k} = (s_{01}\mathbf{s_2}, s_{11}\mathbf{s_2}, \ldots, s_{(p-1)1}\mathbf{s_2}). \tag{7}$$

The Kronecker product is not commutative.

The Chinese product of the two sequences of (6) is defined only when p and q are relatively prime, that is, when they do not have any common prime factor. The Chinese product is defined as

$$s_{Ck} = s_{1i}s_{2j}, \tag{8}$$

where

$$k = \begin{cases} i \bmod p \\ j \bmod q \end{cases} \tag{9}$$

and has the length pq. It is called the Chinese product because the solution of the congruence relation (9) is what the Chinese remainder theorem deals with. It is commutative.

nputationally, the Chinese product of the two sequences can be obtained easily by re-
ting the sequences $\mathbf{s_1}$ and $\mathbf{s_2}$ of relatively prime lengths p and q respectively, q and p
es and then taking an element-wise (Schur) product (Moharir 1992).
he autocorrelation of $\mathbf{s_k}$ can be expressed in terms of the autocorrelations of $\mathbf{s_1}$ and $\mathbf{s_2}$
ollows (Turyn 1968).

eorem 1.

$$\rho_{s_k}(qk_1 + k_2) = \rho_{s_1}(k_1)\rho_{s_2}(k_2) + \rho_{s_1}(k_1 + 1)\rho_{s_2}(q - k_2),$$
$$k_1 = 0, 1, \ldots, p - 1; \quad k_2 = 0, 1, \ldots, q - 1. \tag{10}$$

*he theorem shows that if the individual autocorrelations are good, so is the resultant
ocorrelation, except at some lag values. It can further be shown that the discrimination
he Kronecker product of two sequences depends only on the discriminations of the
mponent sequences. The attenuated minimum guarantee theorem below is important.*

eorem 2. *If the discriminations of the two sequences $\mathbf{s_1}$ and $\mathbf{s_2}$ are D_1, and D_2 respec-
ly, with $\min(D_1, D_2) = D_{\min}$ and $\max(D_1, D_2) = D_{\max}$, then the discrimination D_K
oth their Kronecker products $\mathbf{s_1} \times \mathbf{s_2}$ and $\mathbf{s_2} \times \mathbf{s_1}$ is bounded as*

$$D_{\min} \geq D_K \geq \alpha D_{\min}, \quad \alpha = (D_{\max}/(1 + D_{\max})), \tag{11}$$

ere α may be called the attenuator.

The following has been shown by Moharir (1977).

eorem 3. *The periodic autocorrelation of the Chinese product of two sequences is a
inese product of their periodic autocorrelations.*

A sequence obtained by a bi-parental product (Moharir 1992) of two or more sequences
later be further improved by a Hamming scan. The time requirement comes down
nsiderably because the algorithm begins with a good starting point. The importance of
nlocal optimizers such as the genetic algorithm is that the end result depends on the
rting point only weakly, and ideally, not at all. But this statement is concerned with
y the ultimate reachability of the algorithm, and the utility of a good starting point in
ermining the time requirements should not be underestimated. This is particularly so
en good starting points can be designed or devised. Second, global optimization also is a
al and not a reality. In any case, the Hamming scan is not a global optimization algorithm.
erefore, if a large starting discrimination can be obtained by a simple procedure, in
lition to the time saved, many lower local optima of which the Hamming scan possibly
ld not come out would also be avoided.
The anabolic bi-parental product scheme really works well as can be seen from table 1.
nalyses the role of the Kronecker and the Chinese products in the proposed algorithm.
me ternary sequences with very good discrimination were chosen. They were obtained
m the ternary sequences with the best merit factors listed by Golay (1977) by recursive
mming scan to improve the discriminations. Then their Kronecker and Chinese products

Table 1. Analysis of the efficacy of the bi-parental products, viz. Kronecker and Chinese, in designing ternary sequences with large discrimination. The lengths of the component sequences are n_1 and n_2. Their discriminations are D_1 and D_2. The discrimination of the product sequence is D. The bi-parental product efficiency is η and the exponent is γ. Whether the bi-parental product P is Kronecker or Chinese is indicated by K and C respectively.

n_1	n_2	n	D_1	D_2	P	D	η	γ
31	31	961	13.0000	13.0000	K	12.7547	0.9811	0.4963
9	33	297	7.0000	12.5000	K	7.0000	0.7483	0.4352
33	33	1089	12.5000	12.5000	K	12.5000	1.0000	0.5000
24	111	2664	20.0000	20.0000	K	19.7531	0.9877	0.4979
24	147	3528	20.0000	20.0000	K	19.6721	0.9836	0.4972
24	159	3816	20.0000	21.8333	K	20.0000	0.9571	0.4928
11	13	143	11.0000	13.0000	K	11.0000	0.9199	0.4832
					C	8.9375	0.7474	0.4413
13	15	195	13.0000	13.0000	K	13.0000	1.0000	0.5000
					C	12.0714	0.9286	0.4856
11	17	187	11.0000	11.0000	K	11.0000	1.0000	0.5000
					C	6.7222	0.6111	0.3973
11	23	253	11.0000	19.0000	K	11.0000	0.7609	0.4488
					C	7.7407	0.5354	0.3831
19	31	589	13.0000	13.0000	K	13.0000	1.0000	0.5000
					C	12.0714	0.9286	0.4856
13	33	429	13.0000	12.5000	K	12.5000	0.9806	0.4961
					C	10.1563	0.7967	0.4554
29	33	957	8.6667	12.5000	K	8.6667	0.8327	0.4609
					C	8.5526	1.2141	0.5381
24	145	3480	20.0000	23.2000	K	20.0000	0.9285	0.4879
					C	11.2621	0.5228	0.3944
24	155	3720	20.0000	22.1667	K	20.0000	0.9499	0.4916
					C	19.4161	0.9221	0.4867
15	17	255	13.0000	11.0000	K	11.0000	0.9199	0.4832
					C	11.0000	0.9199	0.4832
11	25	275	11.0000	10.5000	K	10.5000	0.9770	0.4951
					C	10.5000	0.9770	0.4951
13	31	403	13.0000	13.0000	K	13.0000	1.0000	0.5000
					C	13.0000	1.0000	0.5000
9	11	99	7.0000	11.0000	K	7.0000	0.7977	0.4480
					C	8.5556	0.9750	0.4942
9	23	207	7.0000	19.0000	K	7.0000	0.6070	0.3979
					C	8.3125	0.7208	0.4330

(Continued on facing page)

of these two products are listed. It càn be seen that impressive starting points for the Hamming scan are available. The purpose of table 1 is, however, different. It compares the two products as the bases of the bi-parental product signal design. In table 1, six situations can be identified. In view of theorem 2, there is not much uncertainty about what the Kronecker product can achieve. For some entries in table 1, the component lengths are such that only the Kronecker product is defined and the Chinese product is not defined as the two lengths have a common factor. For these sets of component lengths, only the Kronecker product can be used. For the next set of entries, D_C is less than D_K. For the next four sets of entries, D_C equals D_K, lies between D_K and D_{max}, equals D_{max}, and exceeds

le 1. (*Continued*)

n_2	n	D_1	D_2	P	D	η	γ
29	261	7.0000	8.6667	K	7.0000	0.8987	0.4740
				C	8.2727	1.0621	0.5147
29	377	13.0000	8.6667	K	8.6667	0.8165	0.4571
				C	11.2667	1.0614	0.5126
29	435	13.0000	8.6667	K	8.6667	0.8165	0.4571
				C	11.6552	1.0980	0.5198
29	725	10.5000	8.6667	K	8.6667	0.9085	0.4787
				C	9.1000	0.9539	0.4895
25	425	11.0000	10.5000	K	10.5000	0.9770	0.4951
				C	11.0000	1.0235	0.5049
25	525	8.5000	10.5000	K	8.5000	0.8997	0.4765
				C	10.5000	1.1114	0.5235
31	651	8.5000	13.0000	K	8.5000	0.8086	0.4548
				C	13.0000	1.2367	0.5452
21	357	11.0000	8.5000	K	8.5000	0.8790	0.4716
				C	13.3571	1.3814	0.5712
21	399	13.0000	8.5000	K	8.5000	0.8086	0.4548
				C	13.8125	1.3140	0.5580
27	459	11.0000	8.3333	K	8.3333	0.8704	0.4693
				C	11.4583	1.1968	0.5398
29	493	11.0000	8.6667	K	8.6667	0.8876	0.4738
				C	13.0000	1.3314	0.5628
29	551	13.0000	8.6667	K	8.6667	0.8165	0.4571
				C	14.0833	1.3268	0.5599
29	609	8.5000	8.6667	K	8.5000	0.9903	0.4977
				C	9.2083	1.0729	0.5164
29	783	8.3333	8.6667	K	8.3333	0.9806	0.4954
				C	9.4203	1.1085	0.5241
31	465	13.0000	13.0000	K	13.0000	1.0000	0.5000
				C	13.5200	1.0400	0.5076
33	825	10.5000	12.5000	K	10.5000	0.9165	0.4821
				C	13.1250	1.1456	0.5279
33	1023	12.5000	12.5000	K	12.5000	0.9806	0.4961
				C	15.4762	1.2141	0.5381

ax. Thus, the Chinese product has the possibilities of providing better starting points
n those provided by the Kronecker product, for the bi-parental product algorithm for
nal design.

t is useful to define some quantitative measures for the bi-parental products in the
text of the objective function chosen. The Kronecker efficiency η_K (for discrimination)
lefined as

$$\eta_K = D_K/(D_1 D_2)^{1/2}. \tag{12}$$

e Kronecker exponent γ_K is defined by

$$D_K = (D_1 D_2)^{\gamma_k} \quad \text{or} \quad \gamma_K = \log D_K/\log(D_1 D_2). \tag{13}$$

Table 1 shows values of η_K and γ_K. It can be seen that they lie in narrower ranges than is
gested by theorem 2. A starting sequence of a particular large length can be obtained by

factoring the length differently. In such cases two thumb rules could be used for guidance. the Chinese efficiency η_C and the Chinese exponent γ_C can be defined by replacing D_K by D_C in (10) and (11), where D_C is the discrimination obtained by using a Chinese product. Table 1 also shows η_K and γ_K. The ranges of these are larger than those of η_K and γ_K. The extension of the ranges on the higher sides must be viewed as advantageous.

2.3 *Hamming scan and Simon's principle*

Our interest was in obtaining ternary sequences with good discrimination values. The Kronecker product is not commutative. But the discriminations of $s_1 \times s_2$ and $s_2 \times s_1$ are the same. The Chinese product may have a superior, equal or inferior discrimination. Subsequently, however, they may evolve differently under a Hamming scan, so that the initial advantage of a superior discrimination may not last. When starting sequences of a particular length can be obtained by bi-parental products using different factors, further progress under a Hamming scan is also different. These are indicators that a Hamming scan does not give the global optimum. Let the ratio of the discrimination eventually obtained by recursively using a Hamming scan to the discrimination of the starting sequence obtained by a bi-parental product be called the Hamming gain for that product. It is not as constant as the bi-parental product efficiency or the exponent. Let both the components in the Kronecker product be Barker sequence of length 11. Then the discrimination of the resultant sequence of length 121 cannot be improved by the Hamming scan, meaning that the Hamming gain is just unity. But such situations are rather rare. Variability of the Hamming gain makes it difficult to use a good starting point as a very dependable criterion and offers no good thumb rules. One may make a choice on the basis of the average size of the improvement in the first few Hamming scans. But that is more a way of trying to minimize the computational effort than trying to estimate beforehand what discriminations may be achievable, as the improvement is not uniform over successive Hamming scans and the number of successful scans is also not known beforehand. It has been observed that the Hamming gain for the Chinese product is frequently less than that for the Kronecker product. It may also be noted that a sequence obtained by Hamming scan operating on a bi-parental product of two sequences, can later be used as a component in the same or different bi-parental product. All these possibilities have not been exhausted.

However, one more idea has been found to be very useful. After taking a bi-parental product of two good sequences, if the merit factor is improved by the recursive Hamming scan for some time and then the discrimination is taken to be the objective function to be

Table 2. The number of lengths N_T at which ternary sequences exceeding the discrimination threshold of D_T have been obtained.

D_T	N_T	D_T	N_T	D_T	N_T
48	2	44	7	40	27
36	47	34	75	32	119
30	170	28	242	26	304
24	344	22	267	20	380

le 3.   A list of ternary sequences, with discrimination values greater than 36, in the descending order
le discrimination.

gth	Discrimi-nation	Length	Discrimi-nation	Length	Discrimi-nation	Length	Discrimi-nation
4	48.9048	924	42.4667	1584	40.2414	1001	37.6087
1	48.2222	649	42.4615	590	40.0000	592	37.5455
3	47.3125	975	41.7059	582	40.0000	555	37.4000
1	44.6667	1080	41.3889	584	39.8000	596	37.2727
9	44.4444	819	41.1538	1650	39.7667	593	36.9091
2	44.2400	1188	40.9474	1680	39.6000	851	36.8125
1	44.0588	1305	40.8889	1088	38.5714	1121	36.5789
0	43.5385	726	40.8571	925	38.5625	1608	36.5556
0	43.3214	623	40.7000	625	38.1818	1617	36.1389
7	43.1579	858	40.6471	1071	37.9524	1053	36.1200
8	43.0000	1750	40.6333	1173	37.7500	767	36.0000
9	42.5294	792	40.3846	1464	37.6429		

reased by the same procedure, superior discrimination values frequently result, than if
 discrimination was increased from the start.

Results

ly the sequence having the largest discrimination obtained at any length has been re-
led. The numbers of lengths at which various discrimination thresholds have been
ceeded are shown in table 2. Thus the discrimination thresholds of 32, 26 and 18 have
en exceeded at more than 100, 300 and 400 lengths respectively. The details about the
quences having the discrimination values of greater than 36 are shown in table 3.
It is seen that as the length increases it is easier to reach higher discrimination values.
is is a simple consequence of theorem 2. To see this point, assume that the work of
taining ternary sequences with good discrimination values is conducted up to the length
160. Then, the best sequences are of lengths 24, 111, 145, 147, 155 and 159, and have
criminations of 20, 20, 23, 20, 22.1667 and 21.8333 respectively. That is, they all have
criminations exceeding 20. Then, theorem 2 implies that a Kronecker product of any two
them must have a discrimination exceeding 19.05. In general, the actual discrimination
hieved is frequently much better than αD_{min}, as can be seen from table 1. Hamming
in can then raise the discrimination even further. It is very rare that the Hamming gain
ust 1. Thus, the highest discrimination obtained up to any length can almost always be
ceeded at some higher lengths.

Conclusion

e bi-parental product algorithm has given ternary sequences with very good discrimi-
ions. The success of the algorithm indicates that the locally complete search may be
eferable to the Monte Carlo search which depends rather excessively on the efficacy of
ance in the matter of optimization. Whereas chance is a useful ally while dealing with
mbinatorially complex optimization problems, where feasible, it should be helped by

design procedures (Singh *et al* 1996) which can minimise the search effort. The anabolic role of bi-parental products which permitted going to large lengths quickly is particularly noteworthy in this regard. The algorithm has taken the problem of aperiodic signal design one step closer to that of the periodic signal design. The latter has two features. One of them is the availability of the regular construction procedures. That goal is still far away for the former problem. The second feature is that the discovery of a new good sequence means automatic construction of good sequences at many larger lengths. The bi-parental product procedure in the algorithm gives that feature to the aperiodic signal design problem also, though in a weaker sense, as the Hamming scan, which includes choice, is still to be performed.

To the best of our knowledge, the ternary sequences obtained here are the best. Yet, the bi-parental product algorithm is not a global optimisation algorithm. Therefore, there could be procedures which can improve upon the sequences obtained here.

References

Barker R H 1953 Group synchronization of binary digital systems. In *Communication theory* (ed.) W Jackson (London: Butterworths)

Baumert L D 1971 *Cyclic difference sets* (Berlin: Springer-Verlag)

Beenker G F M, Claasen T A C M, Heime P W C 1985 Binary sequences with a maximally flat amplitude spectrum. *Phillips J. Res.* 40: 289–304

Bernasconi J 1987 Low autocorrelation binary sequences: statistical mechanics and configuration space analysis. *J. Phys.* 48: 559–567

Bernasconi J 1988 Optimization problems and statistical mechanics. *Proc. Workshop on Chaos and Complexity 1987* (Torino: World Scientific)

Boehmer M A 1967 Binary pulse compression codes. *IEEE Trans. Inf. Theory* IT-13: 156–167

Brewere J W 1978 Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits Syst.* CAS-25: 772–781

De Groot C, Wurtz D, Hoffman K H 1992 Low autocorrelation binary sequences: exact enumeration and optimization by evolutionary strategies. *Optimization* 23: 369–384

Golay M J E 1972 A class of finite binary sequences with alternate autocorrelation values equal to zero. *IEEE Trans. Inf. Theory* IT-18: 449–450

Golay M J E 1977 Sieves for low autocorrelation binary sequences. *IEEE Trans. Inf. Theory* IT-23: 43–51

Golay M J E 1982 The merit factor of long low autocorrelation binary sequences. *IEEE Trans. Inf. Theory* IT-28: 543–549

Golay M J E 1983 The merit factor of Legendré sequences. *IEEE Trans. Inf. Theory* IT-29: 934–936

Golay M J E, Harris D 1990 A new search for skew-symmetric binary sequences with optimal merit factors. *IEEE Trans. Inf. Theory* 36: 1163–1166

Hoholdt T, Jensen H E, Justesen J 1985 Aperiodic correlations and the merit factor of a class of binary sequences. *IEEE Trans. Inf. Theory* IT-31: 549–552

ıoldt T, Justesen J 1988 Determination of the merit factor of Legendré sequences. *IEEE Trans. Inf. Theory* IT-34: 161–164

land J H 1992 Genetic algorithms. *Sci. Am.* 267: 66–72

sen J M, Jensen H E, Hoholdt T 1991 The merit factor of binary sequences related to difference sets. *IEEE Trans. Inf. Theory* 37: 617–626

dock A M, Meyer R, Bass D 1986 Longest binary pulse compression codes with given peak sidelobe levels. *Proc. IEEE* 74: 366

estler A 1969 Beyond atomism and holism : the concept of the holon. In *Beyond reductionism* (eds) A Koestler, J R Smythies (London: Hutchinson)

halewicz Z 1992 *Genetic algorithms + data structures = evolution programs* (Berlin: Springer-Verlag) p 250

harir P S 1974 Ternary Barker codes. *Electron. Lett.* 10: 460–461

harir P S 1975 Generation of the approximation to binary white noise. *J. Inst. Electron. Telecommun. Eng.* 21: 5–7

harir P S 1976 Signal design. *Int. J. Electron.* 41: 381–398

harir P S 1977 Chinese product theorem for generalized PN sequences. *Electron. Lett.* 13: 121–122

harir P S 1992 *Pattern-recognition transforms* (Taunton: Research Studies Press) p 256

harir P S, Varma S K, Venkatrao K 1985 Ternary pulse compression sequences. *J. Inst. Electron Telecommun. Eng.* 31: 33–40

wmann D J, Byrnes J S 1990 The L norm of a polynomial with coefficients +1. *Am. Math. Monthly* 97: 42–45

ldy V U, Rao K V 1986 Biphase sequence generation with low sidelobe autocorrelation function. *IEEE Trans. Aerosp. Electron. Syst.* AES-22: 128–133 ·

ıon H A 1981 *The sciences of the artificial* (Cambridge, MA: MIT Press)

gh R, Moharir P S, Maru V M 1996 Eugenic algorithm-based search for ternary pulse compression sequences. *J. Inst. Electron. Telecommun. Eng.* 42: 11–19

·yn R 1963 Optimum code study. Sylvania Electric Systems, Rep. F 437-1

·yn R 1968 Sequences with small correlation. In *Error correcting codes* (ed.) H B Mann (New York: Wiley) pp 195–228

rror boxes and mirror mounts for photophysics beamline

P MEENAKSHI RAJA RAO [1], B N RAJASEKHAR [1], N C DAS [1], H A KHAN [1], S S BHATTACHARYA [1], A S RAJA RAO [2] and A P ROY [1]

[1] Spectroscopy Division, Bhabha Atomic Research Centre, Mumbai 400 085, India
[2] Centre For Advanced Technology, Indore 452 013, India

Abstract. A medium resolution beamline, viz. photophysics beamline is being built at INDUS-I, a 450 Mev synchrotron radiation source (SRS) at the Centre for Advanced Technology (CAT), Indore. To house the pre- and post-focusing toroidal mirrors and permit precision movements for steering the SRS beam up to the sample, two ultra-high vacuum (UHV) compatible mirror boxes and mirrors have been designed, fabricated and tested. Details of the setup, including UHV testing ($P < 10^{-9}$ mbar) and performance evaluation of the mirror mounts under UHV conditions, are discussed in this paper.

Keywords. Mirror box; mirror mount; photophysics beamline; residual gas spectrum; INDUS-1.

Introduction

US-I is a 450 Mev synchrotron radiation source (SRS) being constructed at the Centre Advanced Technology, Indore (Ramamurthi 1992). The radiation emitted from the age ring is useful for performing experiments in the soft X-ray to infrared region. eamline, viz. photophysics beamline dedicated to experiments on photophysics at US-I, is under fabrication. This beamline makes use of a toroidal mirror for focusing ight [40 mrad(H) $\times$ 6 mrad(V)] from the tangent point of INDUS-I onto the entrance of a one-metre Seya-Namioka type of monochromator and a second toroidal mirror for sing the monochromatic light emerging from the exit slit of the monochromator onto sample (1 mm $\times$ 1 mm spot size) positioned at a distance of one metre from the centre he focusing mirror. Detailed optical and mechanical layouts of the beamline formed basis of the mirror box and mirror mount designs (Das & Rajasekhar 1992; Meenakshi Rao *et al* 1992). A side view of the photophysics beamline is shown in figure 1. mate pressure in the beamline has to be maintained at $<10^{-9}$ mbar*. In addition to

bar $= 10^2$ Pa
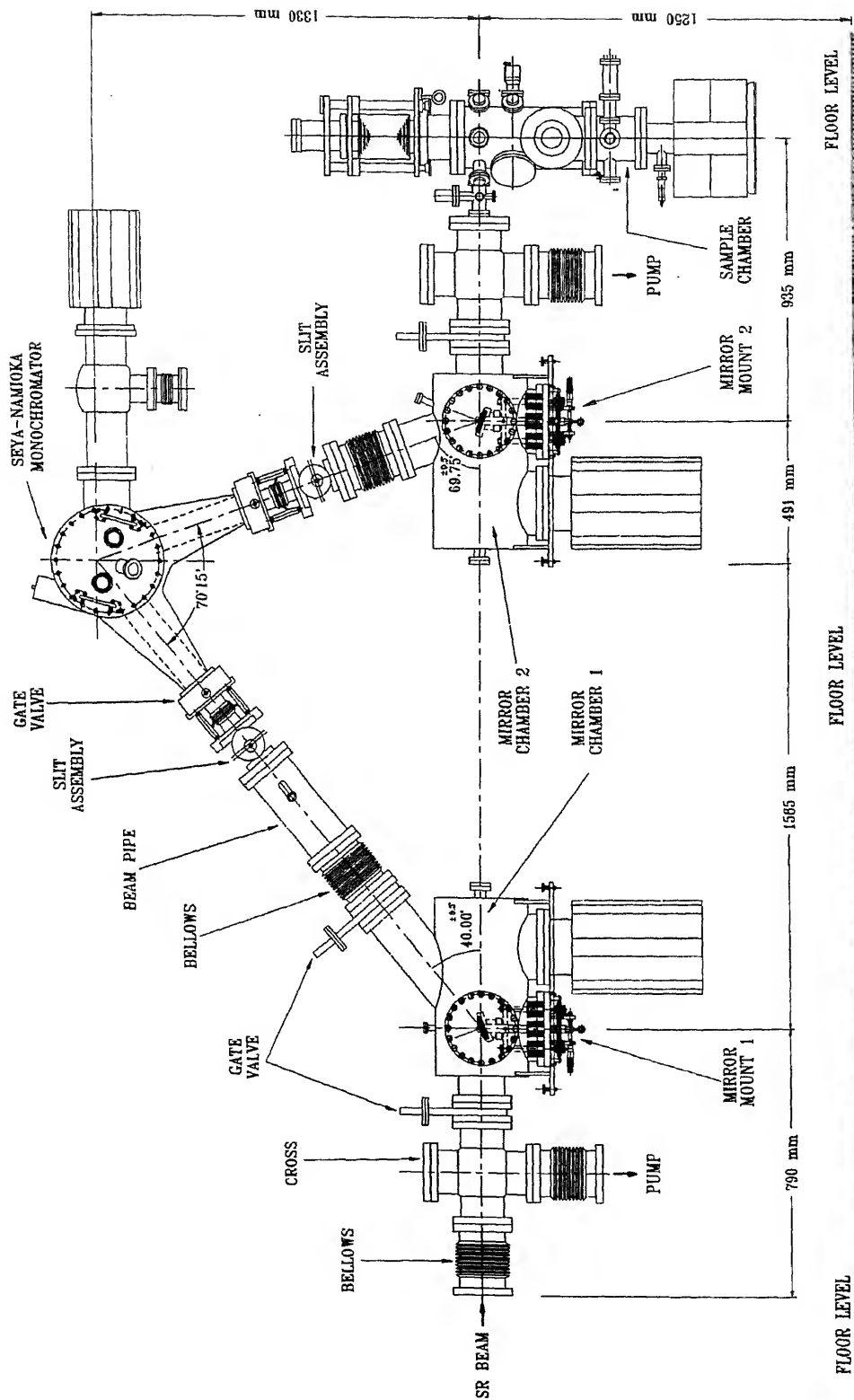
Figure 1. Mechanical layout of photophysics beamline.

g the prime requirement for connecting the beamline to the storage ring which operates ressure of $<10^{-9}$ mbar such ultra-high vacuum (UHV) environment also protects the ces of the optical components. Installation of the beamline to get maximum flux at the le position requires precise and reproducible rotational and translational movements e mirrors about the X, Y and Z axes under ultra-high vacuum conditions so as to the synchrotron radiation source (SRS) beam up to the sample. The fine movements red are achieved with the help of UHV compatible mirror mounts. Details of design, cation and testing of the mirror mounts and mirror chambers are described below.

Description

Design considerations

Mirror mounts: In case of mirror mounts for holding both pre- and post-mirrors, djusting their alignment and fine movements, the considerations are

otational and translational motions in three mutually perpendicular directions (X, Y nd Z) should be independent. They should not disturb the prevailing UHV conditions f the beamline.

ccuracies of translational movements should be of the order of 1 ± 0.5 mm.

ccuracies of rotational movements should be of the order of 1 ± 0.2 milliradian.

the linear and angular positions are not maintained within the accuracies mentioned e, the image at sample position will be shifted from the mean position by about 0.3 mm ting in loss of flux of the order of 25%.

Mirror boxes: The UHV mirror boxes needed several ports of various sizes not to connect them to the beamline but also to connect mirror mounts, gauge heads etc. ldition, ports also were provided for handling the mirror mounts and viewing the beam on the mirror. Table 1 gives a brief description of the ports of the mirror boxes.

1. Description of the ports of the mirror boxes.

e	Quantity	Description
F	1	Mirror mount (Port 1)
F	5	Beam entry (Port 2) Beam exit (Port 3) Manual alignment (Ports 4 & 5) Sputter ion pump (Port 6)
F	3	BA Gauge (Port 7) Roughing pump (Port 8) Pirani–Penning Gauge (Port 9)
F	1	View Port (Port 10)

Port 4

Port 8

Port 7

MIRROR BOX

Port 5

Port 9

Port 3

Port 10

Port 2

Port 6

MIRROR MOUNT

Port 1
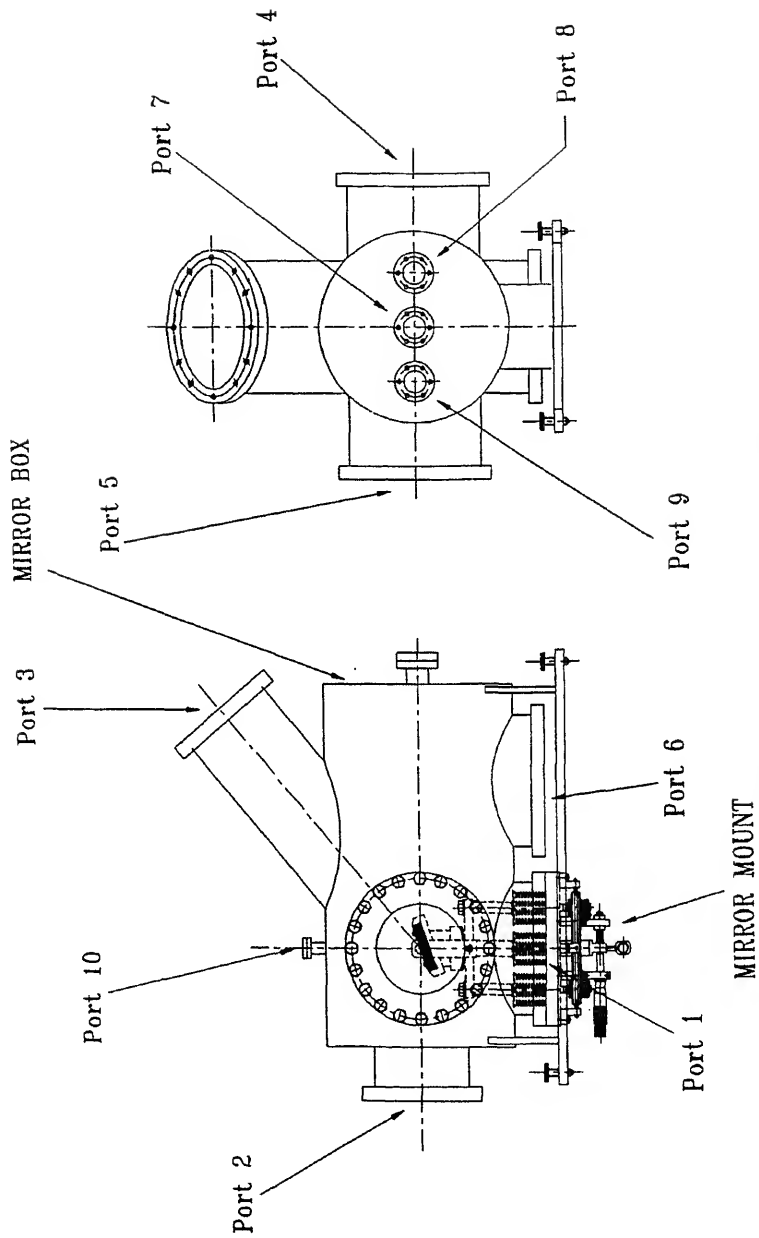
Figure 2.   Schematic of mirror box M_1.
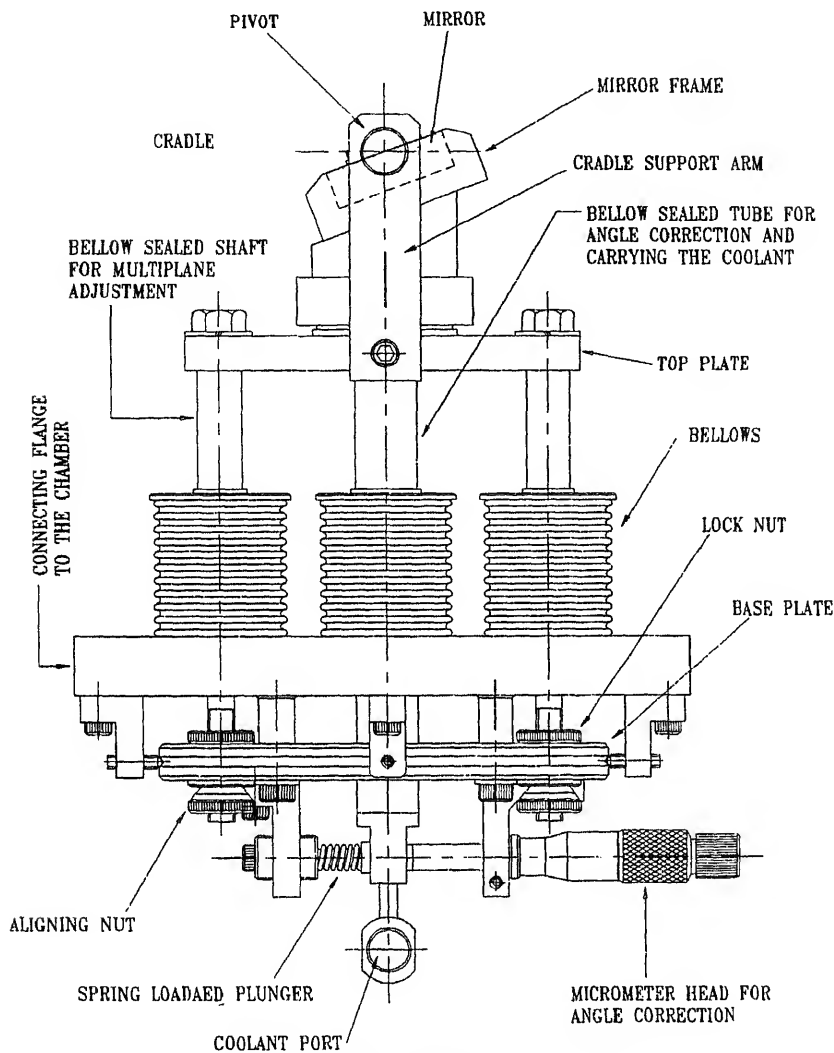
Figure 3. Schematic of mirror mount M_1.
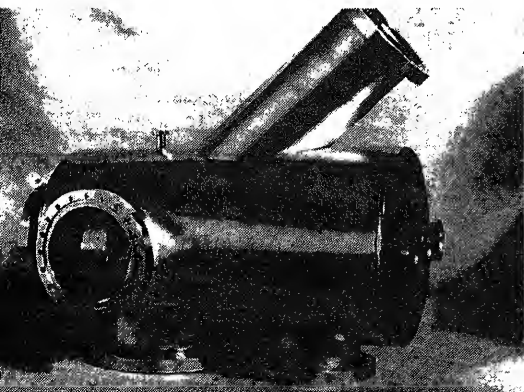
Figure 4. Photograph of mirror box.
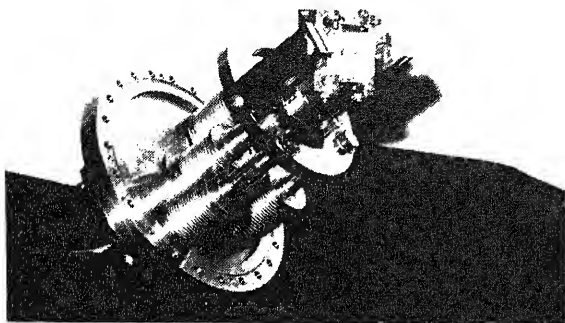
Figure 5. Photograph of mirror mount.

3. Design details

Based on the design considerations mentioned above and on the guidelines from the beam-lines of the Ultraviolet Synchrotron Orbital Radiation Facility (UVSOR), Japan (Sakai *et al* 1987), mirror chambers and mirror mounts were designed and fabricated. The two mirror boxes are mirror images of each other with a few minor differences depending on the entry and exit of light (Ramamurthi 1992; Meenakshi Raja Rao *et al* 1992, 1995). Schematics of the mirror boxes and mirror mounts are shown in figures 2 and 3 respectively. Figures 4 and 5 show photographs of the mirror mount and mirror chamber respectively. The mirror is enclosed in a frame provided with adjustable screws on the mirror mount. This arrangement facilitates the rotation of the mirror around an axis which is perpendicular to the plane of the mirror. The mirror frame is mounted on a copper block having an angle equal to the mirror tilt. The mirror frame has an arrangement for locating the mirror accurately by using fine pitch positioning screws. Two pivot joints provided on the frame impart the rotation necessary for angle correction. Rotation of the mirror is controlled from outside by a bellows-sealed mechanism using a micrometer head and a spring loaded plunger. The tube which imparts this rotation also acts as a carrier for coolant to the mirror frame. Multiplane adjustment of the mirror is provided by using three long bellows-sealed shafts located in a right-angled triangular geometry. The mirror mount is welded to a 203CF flange and this flange is connected to port 1 of the mirror box.

4. Materials and fabrication

SS 304L has been used for fabrication of most of the components of the mirror mounts and mirror boxes. The bellows and coolant channel used are made of SS 316. The supporting block of the mirror holder is made of copper for better thermal conductivity. The chambers and mounts are TIG-welded from inside, cleaned, electropolished, and degassed in a vacuum furnace at a temperature of 800°C for 8 h. This procedure reduced degassing thereby saving time taken for achieving ultimate pressure.

Testing setup

mirror mounts were connected to the mirror chamber after degassing all the compo-
s and the mirror box in a vacuum furnace. A 2701/s sputter ion pump was used to
eve the ultimate pressure. A pirani gauge was used to measure pressure up to 10^{-3} mbar.
ayard–Alpert nude ionisation gauge (model IG5M of M/S Edwards, UK) was used to
sure pressure from 10^{-4} mbar to 10^{-9} mbar and a penning gauge and a leak detector
connected to the chamber with the help of a "T" through a turbomolecular pumping
on (Model Turbopac 5150 of M/S Alcatel, France). The rest of the ports of the cham-
were blanked off. Two sight glasses were provided for the entry and exit of collimated
light from He–Ne Laser used for testing the mirror mount mechanism. All gauges,
ge control units, pumps and their accessories were provided power through the regu-
d power supply. Heating of the chamber, mount, ion pump etc. for degassing purposes
carried out using different heating elements operated through temperature controllers.
the vacuum sealings were done by metal gaskets made of OFHC copper. The experi-
tal setup showing different instruments used for leak detection and evaluation is shown
gures 6 and 7.

1. S.I.P.
2. MIRROR CHAMBER–M1
3. B.A. GAUGE HEAD
4. UHV VALVE
5. T.C.GAUGE HEAD
6. SORPTION PUMP
7. DIAPHRAGM PUMP
8. T.M.P.& ROTARY PUMPING STATION
9. B.A. GAUGE
10. S.I.P. POWER SUPPLY
11. LEAK DETECTOR
12. HEATERS
13. TEMP.CONTROLLER
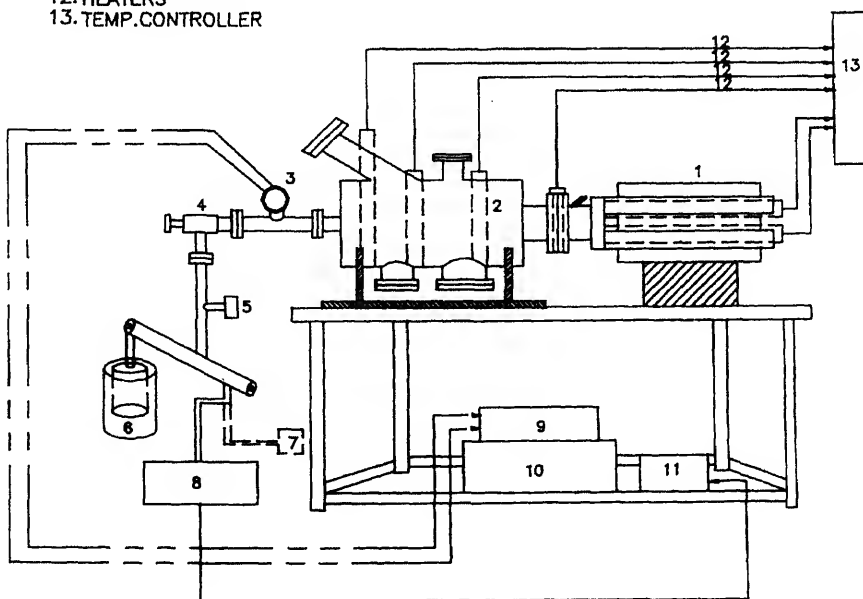
Figure 6. Side view of the testing setup (drawing is not to scale).

1.B.A.GAUGE HEAD
2.UHV ANGLE VALVE
3.THERMOCOPLE GAUGE HEAD
4.BELLOW
5.LEAK DEFECTOR CELL
6.PENNING GAUGE HEAD
7.T.M.P.AND ROTARY PUMP STATION
8.PENNING GAUGE
9.He—LEAK DETECTOR
10.B.A. GAUGE
11.S.I.P.POWER SUPPLY

12.He—CYLINDER
13.ACETONE CONTAINER
14.S.I.P.
15.DIAPHRAGM PUMP
16.SORPTION PUMP
17.MASS ANALYSER HEAD
18.R.F. HEAD
19.RECORDER
20.QMS CONTROLLER
21.PERSONAL COMPUTOR
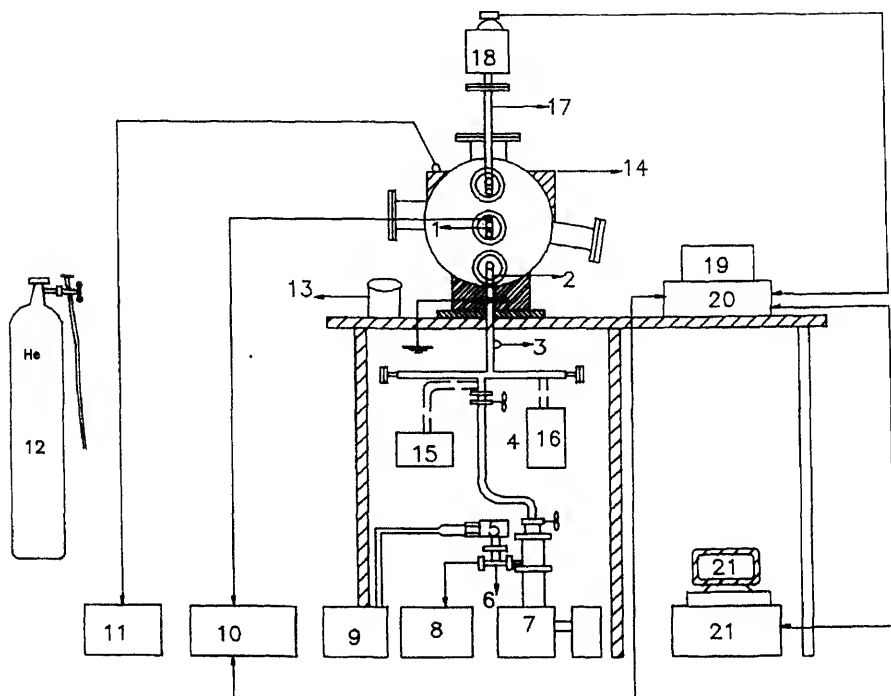
Figure 7. Front view of the testing setup (drawing is not to scale).

6. Testing procedure

To start with, the mirror chamber was leak-tested for welding and gasket leaks of the order of 10^{-10} std.cc/s*. A pressure of 1.0×10^{-8} mbar was achieved with the help of a turbomolecular pump and a sputter ion pump without baking the chamber or the sputter ion pump (unconditioned). The residual gas spectrum was recorded using a quadrupole mass analyzer to ascertain the various components of residual gases and their relative concentrations as shown in figure 8. After baking the sputter ion pump and the chamber for about 8 h at 250°C and pumping the chamber with sputter ion pump, a pressure of 1.0×10^{-9} mbar was achieved. The residual gas spectrum was recorded again at this pressure (figure 9). From figures 8 and 9 it can be seen that water vapour, CO_2, N_2 and hydrocarbons which were the major constituents before degassing, have been suppressed and that H_2 is the only predominant gas component after baking. Thus hydrocarbon-free ultra-high

---

*0.987 std.cc/s $= 0.1$ Pa m^3 s^{-1}
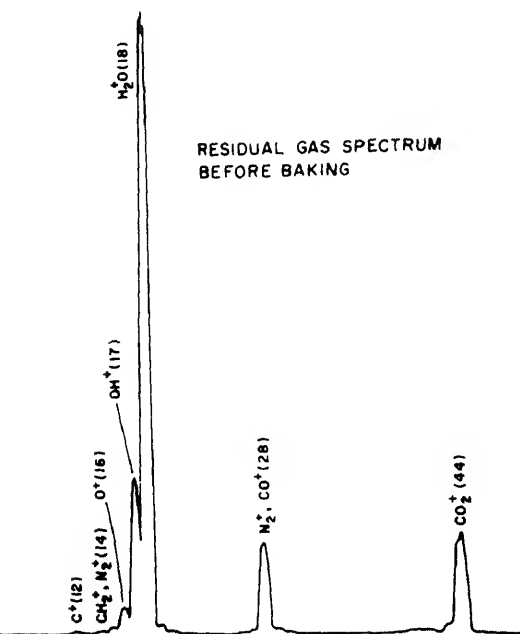
PRESSURE INSIDE THE CHAMBER 4.5 x 10⁻⁸ m.bar
PRESSURE AT THE ANALYSER HEAD 1.0 x 10⁻⁸ m.bar

Figure 8. Residual gas spectrum before baking the mirror chamber.

...um essential for SRS was achieved. After achieving this ultimate pressure, the mirror ...ts were tested for their performance under UHV conditions (i.e., reproducibility ...minimum attainable translational and rotational displacement) using an He–Ne laser. ...test was performed by rotating the mirror about an axis which is perpendicular to ...ncident beam direction and measuring the image displacement on a screen placed ...distance of 1250 mm (corresponding to the distance between pole of the first mirror ...entrance slit of the monochromator). The minimum rotation of the micrometer screw ...rted a rotation of $0.0033°$ to the mirror which resulted in an image displacement of ...at the entrance slit. This is adequate as entrance slit width is planned to be of the order ...0μ. For the second mirror mount similar procedure was repeated and same result of ...is obtained as the minimum displacement of the image and is considered to be very ... for a spot size of $1\,mm \times 1\,mm$ at the sample position.

Conclusions

...mirror chambers and two mirror mounts for photophysics beamline were designed, ...cated and tested. The mirror mounts were tested for their performance at atmospheric ...sure and at a pressure of 1.0×10^{-9} mbar. The residual gas spectrum recorded at ...$\times 10^{-9}$ mbar indicated that H_2 is the major constituent of the residual gases. The ...mum displacement of the image at the entrance slit (for mirror M_1) and at sample (for ...mirror M_2) is 75μ. This value is adequate as slit openings are 100μ for entrance as ...as exit slits and the sample size is $1\,mm \times 1\,mm$.

Figure 9. Residual gas spectrum at ulti mate pressure.

References

Das N C, Rajasekhar B N 1992 Design and evaluation of beamlines on synchrotron source. *Prc Int. Conf. on Synchrotron Radiation Sources*, pp 350–354

Meenakshi Raja Rao P, Das N C, Rajasekhar B N, Kartha V B 1992 Design and development photophysics beamline. *Proc. Int. Conf. on Synchrotron Radiation Sources*, Indore, pp 373–3

Meenakshi Raja Rao P, Rajasekhar B N, Saraswathy P, Das N C, Khan H A, Kartha V B, R Rao A S, Patel R J, Ratnakala K C, Sinha A K, Bhat S 1995 Mirror chambers for photophys beamline at INDUS-I. *Bull. Indian Vacuum Soc.* 26: 7–15

Ramamurthi S S. 1992 Status of INDUS-I and INDUS-II. *Proc. Int. Conf. on Synchrotron Rc ation Sources*, Centre for Advanced Technology, Indore, pp 9–16

Sakai K, Nakamura E, Yamazaki J, Matsudo O, Fukui K, Watanabe M 1987 Cooling of premi at BL1B and BL7B. UVSOR activity report, pp 19–21

An overview of discrete event simulation methodologies and implementation

RAJESH MANSHARAMANI

Tata Research Development & Design Centre, Plot #54B, Hadapsar Industrial Estate, Pune 411 013, India
e-mail: rmansha@pune.tcs.co.in

Abstract. Discrete event simulation has been widely used to model and evaluate computer and engineering systems and has been an on-going area of research and development. This paper presents an overview of the field. It covers specifications of discrete event systems, simulation methodology, simulation languages, data structures for event management, and front and backend support in simulation packages including random number generation and resource management. The emphasis of the survey is on simulation methodology and event scheduling, which forms the core of any simulation package or environment.

Keywords. Discrete event simulation; event scheduling; process interaction; priority queue; simulation languages.

Introduction

The design of a system or a process often needs to be evaluated for correctness and engineering properties before its implementation. Simulation is a cost-effective mechanism to evaluate system and process design. Likewise, to study the behaviour of an existing system or process, simulation is often more cost effective than direct system or process measurement. Depending on the underlying system model, the simulation will take the form of solution of a set of equations, as in the case of a continuous system model, or execution of event-based program code, as in the case of a discrete-event system model. In this paper we will consider only simulation of discrete-event systems.

Ever since the sixties, discrete-event simulation has been widely used for modelling and evaluating computer systems, computer networks, real-time systems, distributed systems, database management systems, manufacturing systems etc. For example, to evaluate the configuration of a computer system for a banking application, to evaluate a resource management policy in an operating system, to study the behaviour of a local area network communication protocol, or to examine strategies for job shop scheduling. A number of

text-books and papers describe traditional as well as non-traditional uses of discrete event simulation, for example, Banks & Carson (1984), Jain (1991), Law & Kelton (1993), Banks & Norman (1996).

Given that several engineering disciplines have found the need for discrete event simulation, several languages and packages have been developed, both for general purpose use, as well as for a suitable set of applications. Advances in software development such as object-oriented design, data structures, and graphical user interfaces have caused advances in simulation techniques and software. As a result, discrete event simulation is an active area of research and development. This paper presents a survey of state of the art in sequential discrete event simulation[1]. The focus will be on simulation methodologies, event scheduling, languages and modelling support in software packages.

The rest of this report is organised as follows. Section 2 defines the terms and formalisms for discrete event systems. Section 3 surveys the different strategies that have been used for discrete event simulation. Section 4 presents a survey of simulation languages and packages. A core aspect of discrete event simulation, that is, future event scheduling is reviewed in § 5. Front-end and back-end support issues are discussed in § 6. Finally, § 7 summarises the main aspects of this paper.

2.   Discrete event systems and their specification

To model any system we first need to define its state space, i.e., the variables that govern the behaviour of the system with respect to the metrics being estimated. If the variables continuously change with time it is called a continuous system. If the system state *instantaneously* changes at discrete points in time, instead of continuously, it is called a discrete system.

In discrete systems whenever some state variables instantaneously change value, this occurrence is denoted by the term *event*. The general behaviour of a discrete event system is that the system starts out with some initial state. The system remains in that state for a duration of time. Then an event occurs which causes the system to instantaneously transit to a new state. This behaviour repeats with the system transiting from state to state over time but remaining in a specific state for a duration of time. This is opposed to continuous systems where the system continuously changes state with time.

Let us consider a simple example of a discrete event system (DEVS), a single server queue that serves customers in first-come-first-serve order. It is assumed that the system is work-conserving, that is, the server will not be idle if there is a customer waiting for service and the server will not abruptly stop serving a customer. We model the state to be the number of customers in the system. The system starts out as empty. The first event to occur is a customer arrival and the state changes to one customer in the system. At this point the server begins service and the customer's departure is scheduled. The next event to occur can either be the second customer's arrival or the first customer's departure, whichever occurs first in time. In the former case, the state will change to two customers

---

[1]Though distributed simulation is an active research topic, the majority of industry uses sequential discrete event simulation on account of the simplicity in description

and in the latter case the state will return to the system being empty. The events can also be specified more elaborately, for example, start service and end service. In this case events occur instantaneously. A customer arrival at an empty queue will cause a start service to occur at the same time, without any change in state[2].

To model a DEVS we need to describe the state space of the system, the events in the system, the state transitions upon events and the times at which events will occur. The event space can be classified further as *external input events, external output events* and *internal events*. External input events are events that are triggered from outside the system, as in the case of a customer arrival at the queue. External output events are those that are generated as output from the system, as in the case of customer departure. Internal events are changes in state variables that do not affect the system environment, e.g., *start service*. The simulation of the discrete event system is done by means of generating events and executing the actions associated with the events.

A DEVS can be precisely modelled using the abstract specification first presented by Zeigler (1976) (see also Evans 1988 and Fishwick 1993). This formalism has been widely used in the DEVS literature. Let $\mathcal{I}$ denote the set of external input event types to the discrete event system. Let $\mathcal{O}$ denote the set of external output event types. Let S denote the set of states of the system, where a subset of the state variables includes the list of future-event times at the given time instant. Then the DEVS abstraction is given by the 7-tuple

$$\langle \mathcal{I}, \mathcal{O}, S, \delta_{int}, \delta_{ext}, \lambda, \tau \rangle,$$

where $\delta_{int} : S \to S$ is the internal transition function dictating state transitions due to internal events.

$\delta_{ext} : \mathcal{Q} \times \mathcal{I} \to S$ is the external transition function dictating state transitions due to external input events. $\mathcal{Q} = \{(s, e) | s \in S, 0 \le e \le \tau(s)\}$ is the total state set of the model; (s, e) represents the state of having been in state s for elapsed time e.

$\lambda : S \to \mathcal{O}$ is the output function generating external events, and

$\tau : S \to \mathcal{R}_0^+$ is the time advance function. If the system is in state s at time t then the system will remain in that state until time $t + \tau(s)$. In other words, $\tau( ) \ge 0$ is the minimum of future event times.

Composite models are constructed in Zeigler (1987) by coupling models (either atomic or composite) by means of *external input coupling, external output coupling*, and *internal coupling*. External input coupling specifies how input events of the composite model are identified with the input events of the components. External output coupling specifies how output events of the composite model are identified with the output events of the components. Internal coupling specifies how the components inside the coupled model are interconnected by means of connections from output events of components to input events of others.

---

[2]If the modeller so desires he can specify the state as the tuple ⟨number of customers in queue, number of customers in server⟩ so that even instantaneous events are associated with a change in state

3. Methodologies of discrete event simulation

Section 2 defined discrete event systems. We now consider how to simulate discrete even
systems, specifically, from a programmer's viewpoint. We describe the most commor
strategies for designing and implementing discrete event simulation programs. (See Evan
1988 for more details.) Three such methodologies are followed in the literature: even
scheduling, activity scanning, and process interaction.

3.1   *Event scheduling*

In this strategy a list of all events in the system is first constructed. Each event is take
individually and described in terms of the particular interaction between entity (say cus
tomer) and resource (say server). Associated with each event is the corresponding actio
or procedure to be invoked when the event occurs. Consider for example a single serve
queue: the events of interest are customer arrival, start service, end service, and leav
system. The processing required at say end service is to compute system metrics, e.g
response time, for the departing customer, clear its resource allocation (if needed), chec
if there is another customer waiting in queue. If so, then schedule a start service event fo
that customer.

   While the approach of event scheduling is straightforward, it involves programming a
a low level. All events have to be enumerated in one place. Care must be taken to ensur
that reactivations are scheduled, as in the case of the start service event being scheduled a
soon as the previous end-service is completed. Likewise, the start service event must als
be scheduled if an arriving customer finds the queue empty. These reactivations typicall
occur in zero time, but if they are not explicitly specified the simulation will terminate
Thus, the responsibility of event scheduling lies entirely on the programmer, which is wh
this strategy is so called. However, because of the explicit specification the efficiency o
this strategy is best as compared to the other strategies.

3.2   *Activity scanning*

The purpose of this strategy is to overcome the reactivation problem of event scheduling
As before, the list of events is first drawn up. But now the events are classified into tw
types: B-activities and C-activities. B-activities are activities that are Bound to occu
whereas C-activities are those that are Conditional. For example, the arrival of a custome
is a B-activity since it is unconditional. However, the start service activity is a C-activit
since it occurs only if the previous event was *end-service* and the queue size is greater tha
zero, or if the previous event was a customer arrival and the queue is empty.

   In this strategy the programmer does not have to explicitly specify reactivations. Th
system automatically handles them. The generic structure of a simulation written in th
strategy consists of three phases. The A-phase advances the time for simulation, the B-pha
checks the type of the B-activity that has occurred at that time and executes the procedu
associated with it, and the C-phase checks the C-activity (or activities), if any, that need
be executed at that time, and executes the corresponding event-handling procedure. No
that the A, B and C phases must proceed in sequential order. Each C-activity has a hea

valuates to true.

While this strategy relieves the programmer from explicitly programming reactivations, he execution is inefficient because all the C-activities need to be checked in the C-phase. If many C-activities are rare-events, then most of the effort spent is wasted. Such inefficiency vas absent in event scheduling because of the explicit invocation of C-activities. To make he processing efficient and yet retain the simulation strategy, attempts have been made to group C-activities and B-activities that share the same entity. Thus, a set of C-activities is grouped with all the B-activities from which they obtain their entities. During the C-phase only those C-activities need to be examined which are grouped with the B-activity that executed.

.3 Process interaction

In the preceding discussion it seems natural to group B and C-activities pertaining to a common entity, and to order them in their actual sequence of succession. Thus, rather than iew the system as a set of event modules it is more natural to view the system as flows of entities.

The process interaction methodology describes the system's workings from the view-oint of an entity flowing through the system. The model is thus described as a projected fe-history of a typical entity, called a process (Evans 1988). To start with, the resources nd entities are identified. Then the entity's resource requirements and interaction with esources, duration of activities etc., are all described in the process.

Each new instance of an entity is a separate process instance. The execution of a process imulation takes a single process instance and executes it until deactivation. That is, if the rocess has to wait for a resource, or if the process is to spend a period of time in a given ctivity (e.g., service time). At that time another process instance is taken and reactivated rom the position where it left off.

Thus, the process oriented approach is modular. All events and activities comprising a rocess are described in the form of the entity model in one place. Note that processes may lso apply to resources. For example, a server process. In the single server queue example, he process oriented simulation can be expressed by means of the entity customer. The equence of description is

    generate customer every inter_arrival
    acquire server (*waits in queue if necessary*)
    get serviced for t time units
    release server
    destroy customer

In this case events are implicitly handled by the language. The modular approach makes it ery easy to concisely specify the customer's behaviour. To implement the process oriented trategy the language needs to provide support for process suspension and reactivation. For xample, when the customer requests for the server and the server is busy then that customer nstance needs to be deactivated. When the server is released then the first waiting customer n the queue needs to be reactivated. This is usually handled by means of a restricted form

of coroutine called *semi-coroutine* (Dahl 1968) or *generator* (Marlin 1980) construct
provided in the simulation language. In this form, the choice of coroutine to be resumed
is determined by a master module that decides on the basis of the next event that is to be
activated.

The process oriented approach allows for high level programming but requires that
the simulation language provide coroutine support, which is typically not the case with
general purpose languages such as C and C++. (See § 4 for extensions to general purpose
languages.) Further, by its very nature it can cause the system to deadlock if two process
instances are waiting for each other to release their resources before acquiring those of
the other. This may not be the intended purpose in cases where more than one resource
is accessed at the same time, but the language does not support atomic access of multiple
resources. (See Evans 1988 for examples.) However, in most cases this is either a non-issue
or special constructs are provided in the language.

3.4   *Summary*

As described in this section, there are three prevalent forms of expressing discrete event
simulation programs: Event scheduling, activity scanning and process interaction. While
the expressive power is the same in each case, they differ in the level of programming.
Event scheduling is the most efficient mechanism in terms of execution speed but involves
explicit specification of events and their actions globally across all parts of the system
as well as reactivations of conditional events. Activity scanning implicitly handles reac-
tivations by explicitly stating which events are unconditional and which are conditional.
However, it incurs a loss of efficiency due to explicit checks for conditional events at time
advances.

Process interaction or process orientation offers a more natural modelling environment
to the programmer since the focus is on the entity and its sequence of activities rather than
on a global perspective of the entire system. It is efficient at run time but requires support for
coroutines in the specification language. With the emergence of distributed systems some
authors have proposed a message based approach to discrete event simulation (Bagrodia
et al 1987) where the physical system is modelled as a set of communication processes.
Events are modelled by message communications. An entity is modelled as a message
communicating process.

4.   Discrete event simulation languages and APIs

Since discrete event simulation is an important field in its own right, a number of languages
have been designed specifically for the purpose. In other cases general purpose languages
have been enhanced. There is also a common trend to provide application programming
interfaces (APIs) for discrete event simulation in a general purpose high level language. In

Historically, discrete event simulation languages were promoted since the 60's. Simscript used the event scheduling strategy, GSP the activity scanning and GPSS the process interaction strategy. Later with the popularity of Algol60, Simula was proposed as a preprocessor to perform simulation programming.

Currently, GPSS is generally regarded as the most popular simulation language. It is a block-structured language that was conceived to run on mainframes and supported by IBM. The GPSS/H processor, a product of Wolverine Software, USA, was first released in the late 1970s and is available on a variety of hardware platforms ranging from PCs to Sun workstations (Scher 1991). The GPSS model is that transactions (i.e., entities) flow through systems to produce dynamic effects. They interact with resources by flowing into *blocks* or program statements. A block remains inactive until a transaction attempts to enter it. Over 40 different blocks describe resource requirements, conditional branching, queueing, data collection, report generation and attribute control (Evans 1988).

Simscript started out as a language for event scheduling but the modern version Simscript II.5 incorporates a process interaction strategy. Like GPSS, Simscript is commercially available (Markowitz *et al* 1987). On the other hand Simula, which is more of an object oriented language, has received less commercial support as a simulation language. SLAM (Pritsker 1986) and SIMAN (Pegden *et al* 1990) have also been used over the last decade. Recently, there is a trend towards object oriented simulation languages as in the case of MODSIM III which is commercially available from CACI, and HSL which was proposed in Sanderson *et al* (1991).

In general, event scheduling languages provide features for specifications of events and associated actions. The actions will be the code necessary to update the system's state and generation of a random time for a future event. The process-oriented languages, typically, include a richer set of language constructs which apart from time delays include interaction of processes with resources, such as suspend at a specific queue, resume once the head of queue is reached, or suspend until is specific condition is satisfied in the form of a *wait-until* construct. The constructs in GPSS, for example, include GENERATE an entity, ENTER into a resource, ADVANCE the time spent at a resource, QUEUE at a resource, LEAVE the resource and TERMINATE an entity.

A number of other languages have been proposed in the literature (see Evans 1988 for example) and many comparisons have been done (cf. Tocher 1965, Dahl 1968 and Virjo 1972). Simulation languages commercially available for personal computers have also been compared, e.g. the comparison between *Simple*$_1$ and *Simian* in Houten (1988).

4.2 *Extensions and APIs*

The primary disadvantage of a special purpose simulation language is that it has to be learned and compilers for it have to be bought. To overcome this problem a number of extensions and application programming interfaces (APIs) for simulation have been proposed for general purpose languages.

Such an effort started with GASP in the 60s. GASP is a collection of FORTRAN subroutines for simulation, and has been very popular among FORTRAN users. Another popular choice has been to extend Pascal for quasi-parallel programming (Kaubisch *et al* 1976; Kriz & Landmayr 1980) or provide APIs in it (Hac 1982; Marsden 1984). Extensions to Algol68 have been proposed in Shearn (1975). Use of PL/I for discrete event simulation has been proposed by Hac (1984), of Ada by Bruno (1984), of Modula-2 in HPSIM (Sharma & Rose 1988) and of SR by Olsson (1990).

In the late 80's and in the 90's, several libraries for discrete event simulation are available in C and C++. In C++ it is possible to write class libraries that support coroutines (cf. Stroustrup & Shapiro 1987) and this allows for process oriented simulation. Process oriented packages in C or C++ include CSIM (Schwetman 1988, 1990), SIM++ (Lomow & Baezner 1990) and YacSim from Rice University. Event scheduling packages include SMPL (MacDougall 1987), and SimPack (Fishwick 1992). More recent packages in C++ provide object-oriented simulation features, for example, C++SIM (Little & McCue 1994) and Awesime (Grunwald 1991).

Apart from extensions and libraries some packages use only diagrams as a means of modelling and simulation. This obviates the need to learn a special purpose language, but it can be rather cumbersome if there are a number of diagrammatic blocks as in GPSS. On the other hand specific tools such as Petri-net analysers (cf. Evans 1988) or queueing network simulators (cf. Melamed & Morris 1985, Funka-Lea *et al* 1991) are useful within the range of applications they model. To model applications beyond their limits one must opt for a programming language. More recently, Shanbagh & Gopinath (1997) have proposed a C++ simulator generator from graphical specifications.

5.   Future event management

No matter what be the simulation strategy or the simulation language, in discrete event simulation the underlying mechanism is scheduling of events. Events are generated for future times. At any time the next scheduling instant is the minimum of the future event times. In practice, the number of future events may range from a handful, say in a single server queueing system at light to moderate load, to a huge number of events, say in the simulation of a wide area network.

The following operations are needed on the *future event list* data structure: *insert* an event (represented by its time and a pointer to associated event information), *delete minimum* time event and return the information pointer, *delete* or *cancel* any arbitrary event from the list. The first two are the most common operations, the last one is used occasionally in some simulation applications, e.g., resource preemption.

The abstract data type appropriate for future event scheduling is the *priority queue* which can be implemented in many ways. The performance of the implementation is subject to the operational profile of *inserts*, *delete-mins*, and *cancels*. Not surprisingly, the

5.1   *Future event list implementations*

The simplest representation of the future event list is an array or linked list ordered by time. Though simple to implement it is rather inefficient for large list sizes since search time is linear in the size of the list. For this reason researchers have proposed index structures over linked lists (Wyman 1976; Franta & Maly 1977, 1978; Henriksen 1977, 1983; Comfort 1979; Nevalainen & Teuhola 1979; Davey & Vaucher 1980; Blackstone *et al* 1981; O'Keefe 1985), a popular scheme among these being that of Henriksen (1977, 1983), which has been incorporated into GPSS (Henriksen & Crain 1982).

Another common approach is to use a *d-heap*. The *2-heap* was proposed by Williams (1964) and then generalised to $d > 2$ by Johnson (1975). In a *d-heap* the nodes are maintained in *heap-order*, where the value of a node is no less than the value of its parent. The *d-heap* is a complete *d-ary* tree satisfying heap order. Using breadth first search the nodes can be indexed into a single array. Other heap-based implementations include the *leftist tree* (Crane 1972, Knuth 1973b), *pagoda* (Francon *et al* 1978), *skew heap* (Sleator & Tarjan 1983, 1986), *binomial queue* (Vuillemin 1978), *pairing heap* (Fredman *et al* 1986, Stasko & Vitter 1987), *Fibonacci heap* (Fredman & Tarjan 1987), *relaxed heap* (Driscoll *et al* 1988) and *radix heap* (Ahuja *et al* 1990).

Search tree-based structures have been popular as well. Binary search trees are the natural choice and have been analysed for future event scheduling in Evans (1983) and Vaucher & Duval (1975). A variant, called *p-tree*, to combine the advantage of linear list and efficiency of tree structures was proposed in Jonassen & Dahl (1975). Among balanced trees or partially balanced trees the most popular version for priority queues is the *splay tree* (Sleator & Tarjan 1985).

A particular type of implementation called *calendar queue* was proposed by Brown (1988) and also independently proposed by Davidson (1989). In this representation time is split into buckets and keys fall within bucket ranges. Indices are wrapped around for the 'next year'. A more recent structure called *fishspear* (Fischer & Paterson 1994) has worst case performance as the d-heap but is oriented towards better performance in the common case and can also be implemented for sequential storage.

5.2   *Analyses and performance comparisons*

Ordinary linked lists sorted by time require $O(n)$ time for insert and $O(1)$ time for delete-min and delete. On the other hand *d-heaps* require $O(\log n)$ for insert, delete and delete-min (Tarjan 1983). Bollobas & Simon (1985) analyse repeated random insertions into a heap where each ordering of the inserted elements is equally likely. They obtain that the number of exchanges per insertion is bounded by a constant of about 1.76. Fibonacci heaps on the other hand have $O(\log n)$ amortised time for delete and delete-min and $O(1)$ for insert (Fredman & Tarjan 1987). Driscoll *et al* (1988) show that these times hold in the worst case for relaxed heaps.

Various comparisons of priority queue implementations have been reported in the literature. Several comparisons have been done under the *hold model* (Vaucher & Duval 1975) where a *hold* operation is one that removes an event from the priority queue and schedules a new event after an interval of time d from a specific distribution $\mathcal{F}$. The hold model

consists of a sequence of hold operations and is parameterised by the number of events in the event queue and the distribution $\mathcal{F}$. It has been used in many of the early studies including Comfort (1979), Davey & Vaucher (1980), Englebrecht-Wiggans & Maxwell (1978), Franta & Maly (1977, 1978), Henriksen (1977), Jonassen & Dahl (1975), Vaucher & Duval (1975), and Vaucher (1977).

Jones (1986) compared several implementations under the hold model and showed many to outperform heaps. The splay tree was shown to have best performance in his study. In a later study Brown (1988) showed that under the hold model the calendar queue performs better than the linear linked list and the splay tree. Chung *et al* (1993) used a Markov hold model to evaluate 14 implementations. Also, using a token ring simulation for comparison they recommend using the splay tree and the calendar queue while stating that heaps are quite 'stable' albeit with lower performance.

McCormack & Sargent (1981) compare several implementations from Comfort (1979), Davey & Vaucher (1980), Franta & Maly (1977), Henriksen (1977), Taneri (1976), Ulrich (1978), Vaucher & Duval (1975) and Wyman (1976), and show that results from real simulation runs are different from that when the hold model is used. They show that Henriksen's method (Henriksen 1977) and the modified heap perform well and are less sensitive to scheduling distributions.

Thus, in general, it is not readily apparent which implementation works best for a given simulation application. Worst case analyses do not reflect performance accurately for the average case. Average case analyses have been done under restrictive assumptions or special cases of applications. It will be desirable therefore for simulation packages to adopt a variety of future event management mechanisms as in SimPack (Fishwick 1992). A knowledgeable user can select the right mechanism but what would be more desirable is a high level interface to select the right mechanism for the simulation's operational profile. Typically, in simulations one needs to run several experiments in the debugging stage itself and during this phase the various priority queue implementations can be compared.

6.   Front end and back end support

Though scheduling of future events forms the core of discrete event simulation, there are a number of other features that are desirable in a simulation environment. They can be classified into front end requirements in the form of diagram editing and graphical output, and back end support in the form of random number generation, resource management, statistical libraries. This section first specifies desirable features of simulation front ends and then desirable features of back ends.

6.1   Simulation front end

The simulation front end must in the least capture the system topology and possibly simplify the model description in terms of user input, and display graphical output of simulation results.

The most general case of user input should allow for a diagram editor to specify user-defined icons to represent processes or resources, connectivity across icons, and connectivity constraints if any. Few packages allow this, however. GPSS has a cumbersome diagram

notation with over 40 different diagram types representing equivalence to program statements. Typically, no package allows for general purpose model specifications. However, specific application domains such as queueing network simulations capture all user specifications through the front end as in the case of PAW (Melamed & Morris 1985) and its successor $Q+$ (Funka-Lea *et al* 1991; also see Shanbagh & Gopinath 1997).

6.2 *Simulation back end*

All simulation environments include support for random number generation, and resource management. Some also include statistical libraries. We elaborate on each of these features below. Note that our emphasis is only on what is provided in standard packages. There are other important aspects such as variance reduction of output and rare-event simulation, which are not covered in this paper[3].

6.2a *Random number generation:* Random number generation forms an integral part of a simulation environment. The underlying model of a discrete event system assumes that the system remains for a given time in each state. This duration of time is modelled using a random number distribution, for example, the inter-arrival time at a queue is often modelled as an exponential distribution. Likewise, service time of a customer in a queue, number of database items that a query will access can be modelled using random number distributions.

Every operating system is usually equipped with a random number generator that generates uniform random variates. In simulation we additionally require generation of non-uniform random variates. Generation of random numbers is more difficult than what one might expect. As Knuth (1973a) says, "Random numbers should not be generated by a method chosen at random."

Several packages and studies have used defective random number generators. For instance, the study of Majumdar *et al* (1988) that simulated performance of parallel processor allocation policies used a defective technique for random number generation which led to incorrect policy comparisons. This was later corrected in Leutenegger & Vernon (1990). The 1988 version of CSIM (Schwetman 1988) used the *rand*( ) function which is well known to have poor random number generation as given in the UNIX system manual page for *random*( ). A survey of more than 50 computer science text books that contained software for random number generation revealed that most of these generators are unsatisfactory (Park & Meller 1988). This shows the importance of using reliable random number generators as given in Knuth (1973a), Park & Meller (1988), and L'Ecuyer (1988).

Any simulation environment must support a variety of distributions, both discrete and continuous. There should be support for multiple random number streams. The package should allow for transformations on random variables to support practical distributions as well as allow for empirical distributions as obtained from measured data. Typical discrete distributions include uniform, Bernoulli, binomial, geometric, Poisson and typical continuous distributions include uniform, exponential, Erlang, hyper-exponential, normal

---

[3]The interested reader can find details in the July 1993 issue of *ACM Transactions on Modelling and Computer Simulation*

numbers for many distributions.

The general techniques that are used for non-uniform random number generation are the *inverse method* and *acceptance-rejection*. In the former, a random number is generated by first generating a uniform number between 0 and 1, and then using it as an argument to the inverse of the distribution. In the latter method, the required density is bounded by that of a scaled version of another distribution for which it is known how to generate random samples. The samples from the known distribution are repeatedly taken until one falls under the required distribution.

6.2b *Resource management:* Discrete event simulation is widely used to study the behaviour of resource contention. For example, contention for CPU and disk in computer systems, contention for database in DBMS, contention for machines in job shops, contention for toll booths on highways, etc. Associated with each resource is a resource handler and *contention queue(s)* to store contending entities. The resource handler decides how to schedule entities from the contention queue(s) on to the resource.

The resource by itself may contain multiple servers, as in the case of a parallel processor or a petrol bunk. The entities may all contend in a single contention queue or may be split across several queues each contending for a subset of the servers. Most simulation environments provide support for single server single queue resources. Some provide support for multiple server single queue resources.

The resource handling discipline can be preemptive or non-preemptive depending on the application in hand. In computer systems preemption is very common at the CPU (but not at disk) whereas in manufacturing systems preemption of executing jobs at plant is typically absent. Among non-preemptive disciplines the most common ones are first come first serve (FCFS) and fixed priority. Some systems also provide support for first fit and best fit. Among preemptive disciplines the most common one is fixed priority with preemptive resume. In CPU scheduling round robin is a common preemptive discipline where preemption occurs on every time quantum.

Having built-in resource scheduling disciplines simplifies the work of the programmer who is now given access to insertion and deletion of entities in contention queues. If the programmer desires to use a very specific discipline, the interface for using the discipline must be the same as that for ones provided by the simulation environment. The simulation environment should provide the facility to integrate custom resource schedulers.

6.2c *Statistical libraries:* The purpose of discrete event simulation is to study the behaviour of a given system. The behaviour of interest to the user is usually captured in the form of metrics such as average and variance of response time, throughput and resource utilisation. To correctly estimate these metrics the programmer needs to insert *measurement probes* at appropriate places in the program. Good simulation environments provide support for probes, that is, creation and initialisation, sampling, determining averages, variance and distributions of accumulated data, as well as confidence intervals. A sophisticated package will provide support for different types of probes, e.g., space average and time average.

Accumulation of samples can vary according to the estimation method being used. Two common techniques for estimating results are regenerative simulation and the method of batch means (Law & Kelton 1993). Regenerative simulation is widely applicable and produces *correct* results (Welch 1983). It essentially estimates metrics at system regeneration points[4] and determines confidence intervals (i.e., estimates of the variance of the metric being analysed) across regeneration points. When enough regeneration points have been encountered to meet the desired confidence interval the simulation stops. While this method produces correct results as per renewal theory, it can be rather costly in large systems to generate regeneration points. For this purpose the method of batch means is a more efficient approach, where metrics are estimated at the end of a given batch size of samples and confidence intervals are calculated across batches. The batch size must be chosen with caution since a small batch size can cause correlation between successive batches. For more details on statistical techniques to estimate steady state behaviour see Pawlikowski (1990).

7. Summary

We have surveyed the field of discrete event simulation on uniprocessors. We have summarised the formal specifications for discrete event systems. Various strategies for simulation, that is, event scheduling, activity scanning and process interaction have been reviewed. Discrete event simulation languages and extensions and APIs of general purpose languages for discrete event simulation have been briefly covered. Data structures for future event management have been surveyed. These include simple linked lists, heaps and variants and a variety of search trees and assorted data structures. Front end and back end support for simulation have been described. Note that topics such as output analysis, variance reduction techniques, rare event simulation are specialised topics that deserve a separate survey in their own right, and have been treated as outside the scope of this paper. Likewise, emerging technologies such as object oriented simulation have not been covered.

Currently, the trend has been to enable the user to build a simulation model of the system under consideration and to efficiently run the simulation code. Not much emphasis has been given on separating modelling from simulation as is prevalent in the continuous simulation world, e.g., simulation of chemical process plants. It would be desirable to create model libraries of resources or of subsystems which can be used for a variety of applications to be simulated. Typically, the approach is to rewrite code from simulation to simulation. This is not only wasteful in terms of development time but also incurs greater chances of bugs in the simulation.

References

Ahuja R K, Melhorn K, Orlin J B, Tarjan R E 1990 Faster algorithms for the shortest path problem. *J. Assoc. Comput. Mach.* 37: 213–223

Bagrodia R L, Chandy K M, Misra J 1987 A message-based approach to discrete event simulation. *IEEE Trans. Software Eng.* 13: 654–665

Banks J, Carson J S 1984 *Discrete-event system simulation* (Englewood Cliffs, NJ: Prentice Hall)

Banks J, Norman V 1996 Second look at simulation software. Non-traditional uses can lead to unexpected benefits. *OR/MS Today* 23: 4

Blackstone J H, Hogg G L, Phillips D T 1981 A two-list synchronization procedure for discrete event simulation. *Commun. ACM* 24: 825–829

Boas P V E, Kaas R, Zijlstra E 1977 Design and implementation of an efficient priority queue. *Math. Syst. Theory* 10: 99–127

Bollobas B, Simon J 1985 Repeated random insertions into a priority queue. *J. Algorithms* 6: 466–477

Brown R 1988 Calendar queues: a fast O(1) priority queue implementation for the simulation event set. *Commun. ACM* 31: 1220–1227

Bruno G 1984 Using Ada for discrete event simulation. *Software Pract. Exper.* 14: 685–695

Chung K, Sang J, Rego V 1993 A performance comparison of event calendar algorithms: an empirical approach. *Software Pract. Exper.* 23: 1107–1138

Comfort J C 1979 A taxonomy and analysis of event set management algorithms for discrete event simulation. In *Proc. 12th Annu. Simulation Symposium,* pp 115–146

Crane C A 1972 Linear lists and priority queues as balanced binary trees. Tech. Rep. STAN-CS-72-259, Comput. Sci., Stanford, CA

Dahl O J 1968 Discrete event simulation languages. In *Programming Languages* (ed.) F Genuys (London: Academic Press)

Davey D, Vaucher J 1980 Self-optimizing partition sequencing sets for discrete event simulation. *INFOR J.* 18: 21–41

Davidson G A 1989 Calendar P's and queues. *Commun. ACM* 32: 1241–1243

Devroye L 1986 *Non-uniform random variate generation* (New York: Springer Verlag)

Driscoll J R, Gabow H N, Shrairman R, Tarjan R E 1988 Relaxed heaps: an alternative to Fibonacci heaps with applications to parallel computation. *Commun. ACM* 31: 1343–1354

Englebrecht-Wiggans R, Maxwell W L 1978 Analysis of the time indexed list procedure for synchronization of discrete event simulations. *Manage. Sci.* 24: 1417–1427

Evans J B 1983 *Investigations into the scheduling of events and modelling of interrupts in discrete event simulation.* Ph D thesis, Dept. of Operations Research, Univ. of Lancaster

Evans J B 1988 *Structures of discrete event simulation: An introduction to the engagement strategy* (Chichester: Ellis Horwood)

Fischer M J, Paterson M S 1994 Fishspear: a priority queue algorithm. *J. Assoc. Comput. Mach.* 41: 3–30

Fishwick P A 1992 SimPack: getting started with simulation programming in *C* and C++. In *Proc. Winter Simulation Conference,* Arlington, VA, pp 154–162

Fishwick P A 1993 A simulation environment for multimodeling. *Discrete Event Dynamic Syst.: Theor. Appl.* 3: 151–171

Francon J, Viennot G, Vuillemin J 1978 Description and analysis of an efficient priority queue

anta W R, Maly K 1977 An efficient data structure for the simulation event set. *Commun. ACM* 20: 596–602

anta W R, Maly K 1978 A comparison of HEAPS and the TL structure for the simulation event set. *Commun. ACM* 21: 873–875

edman M L, Tarjan R E 1987 Fibonacci heaps and their uses in improved network optimization problems. *J. Assoc. Comput. Mach.* 34: 596–615

edman M L, Sedgewick R, Sleator D D, Tarjan R E 1986 The pairing heap: a new form of self-adjusting heap. *Algorithmica* 1: 111–129

nka-Lea C A, Kontogiorgos T D, Morris R J, Rubin L D 1991 Interactive visual modeling for performance. *IEEE Software* 8(5): 58–68

unwald D 1991 A users guide to Awesime: an objected oriented parallel programming and simulation system. Tech. Report CU-CS-552-91, University of Colorado, Boulder.

c A 1982 Computer system simulation in Pascal. *Software Pract. Exper.* 12: 777–784

c A 1984 PL/I as a discrete event simulation tool. *Software Pract. Exper.* 14: 692–702

enriksen J O 1977 An improved events list algorithm. In *Proc. Winter Simulation Conference*, pp 554–557

enriksen J O 1983 Event list management – a tutorial. In *Proc. Winter Simulation Conference*, pp 543–551

enriksen J O, Crain R C 1982 *GPSS/H user's manual* 2nd edn (Annandale: Wolverine Software Corp.)

outen 1988 Simulation languages for PCs take different approaches. *IEEE Software* 5: 91–94

in R 1991 *The art of computer system performance analysis: Techniques for experimental design, measurement, simulation and modelling* (New York: Wiley)

hnson D B 1975 Priority queues with update and finding minimal spanning trees. *Info. Proc. Lett.* 4: 53–57

nassen A, Dahl O J 1975 Analysis of an algorithm for priority queue administration. *BIT* 15: 409–422

nes D W 1986 An empirical comparison of priority queue and event set implementations. *Commun. ACM* 29: 300–311

aubisch W H, Perrott R H, Hoare C A R 1976 Quasiparallel programming. *Software Pract. Exper.* 6: 341–356

nuth D E 1973a *The art of computer programming: Vol. 2/Seminumerical algorithms* (Reading, MA: Addison-Wesley)

nuth D E 1973b *The art of computer programming: Vol. 3/Sorting and searching* (Reading, MA: Addison-Wesley)

riz J, Landmayr H 1980 Extensions of Pascal by coroutines and its application to quasiparallel programming and simulation. *Software Pract. Exper.* 10: 773–789

Ecuyer P 1988 Efficient and portable random number generation. *Commun. ACM* 31: 742–749, 774

aw A M, Kelton W D 1993 *Simulation, modeling and analysis* (New York: McGraw-Hill)

eutenegger S T, Vernon M K 1990 The performance of multiprogrammed multiprocessor scheduling policies. In *Proc. ACM SIGMETRICS* 18: 226–236

ttle M C, McCue D L 1994 Construction and use of a simulation package in C++. *C User's J.* 12: 3

omow G, Baezner D 1990. A tutorial introduction to object-oriented simulation and Sim++. In *Proc. Winter Simulation Conference*, pp 149–153

acDougall M H 1987 *Simulating computer systems: techniques and tools* (Boston: MIT Press)

Majumdar S, Eager D, Bunt R 1988 Scheduling in multiprogrammed parallel systems. In *Proc. ACM SIGMETRICS* 16: 104–113

Markowitz H M, Kiviat P J, Villaneuva R 1987 *Simscript II.5 programming language* (Los Angeles: CACI)

Marlin 1980 Coroutines. In *Lecture notes in computer science 95* (Berlin: Springer-Verlag)

Marsden B W 1984 A standard pascal event simulation package. *Software Pract. Exper.* 14: 659–684

McCormack W M, Sargent R G 1981 Analysis of future event set algorithms for discrete event simulation. *Commun. ACM* 24: 801–812

Melamed B, Morris R J 1985 Visual simulation: the performance analysis workstation. *IEEE Comput.* 18: 87–94

Nevalainen O, Teuhola J 1979 Priority queue administration by sublist index. *Comput. J.* 22: 220–225

Olsson R A 1990 Using SR for discrete event simulation. *Software Pract. Exper.* 20: 1187–1208

O'Keefe R M 1985 Comment on "Complexity Analysis of Event Set Algorithms". *Comput. J.* 28: 245–272

Park S K, Meller K W 1988 Random number generators: good ones are hard to find. *Commun. ACM* 31: 1192–1201

Pawlikowski K 1990 Steady state simulation of queueing processes: a survey of problems and solutions. *ACM Comput. Surv.* 22: 123–170

Pegden C D, Sadowski R P, Shannon R E 1990 *Introduction to simulation using SIMAN* (Sewickley: System Modeling)

Pritsker A A 1986 *Introduction to simulation and SLAM II* (New York: Halstead)

Sanderson P, Sharma R, Rozin R, Treu S 1991 The hierarchical simulation language HSL: a versatile tool for process-oriented simulation. *ACM Trans. Modeling Comput. Simulation* 1: 113–153

Scher J M 1991 Reworked GPSS/H book is a strong standard. *IEEE Software* 8(4): 105–106

Schwetman H 1988 Using CSIM to model complex systems. In *Proc. Winter Simulation Conference,* pp 246–253

Schwetman H 1990 Introduction to process-oriented simulation and CSIM. In *Proc. Winter Simulation Conference,* pp 154–157

Shanbagh V K, Gopinath K 1997 A C++ generator from graphical specifications. *Software Pract. Exper.* 27: 395–424

Sharma R, Rose L L 1988 Modular design for simulation. *Software Pract. Exper.* 18: 945–966

Shearn D C 1975 Discrete event simulation in ALGOL68. *Software Pract. Exper.* 5: 279–293

Sleator D D, Tarjan R E 1983 Self-adjusting binary trees. In *Proc. ACM SIGACT Symp. on Theory of Computing,* pp 235–245

Sleator D D, Tarjan R E 1985 Self-adjusting binary search trees. *J. Assoc. Comput. Mach.* 32: 652–686

Sleator D D, Tarjan R E 1986 Self-adjusting heaps. *SIAM J. Comput.* 15: 52–69

Srikanth S 1996 A software tool for performance analysis of data structure representations. M Tech thesis, Dept. of Comput. Sci. & Eng., Regional Engineering College, Warangal

Stasko J T, Vitter J S 1987 Pairing heaps: experiments and analysis. *Commun. ACM* 30: 234–249

Stroustrup B, Shapiro J E 1987 A set of C++ classes for co-routine style programming. In *Proc. USENIX C++ Workshop,* pp 417–439

Taneri D 1976 The use of subcalendars in event driven simulations. In *Proc. Summer Simulation Conference,* pp 63–66

Tarjan R E 1983 *Data structures and network algorithms* (Philadelphia: SIAM)

Tocher K D 1965 Review of simulation languages. *Oper. Res. Q.* 16: 189–217

Ulrich E G 1978 Event manipulation for discrete simulations requiring large numbers of events. *Commun. ACM* 21: 777–785

Vaucher J G 1977 On the distribution of event times for the notices in a simulation event list. *INFOR J.* 15: 171–182

Vaucher J G, Duval P A 1975 A comparison of simulation event list algorithms. *Commun. ACM* 18: 223–230

Virjo A 1972 A comparative study of some discrete-event simulation languages. In *Proc. Nordata Conference,* Helsinki, pp 1532–1564

Vuillemin J 1978 A data structure for manipulating priority queues. *Commun. ACM* 21: 309–314

Welch P 1983 Statistical analysis of simulation results. In *Computer performance modeling handbook* (ed.) S S Lavenberg (New York: Academic Press)

Williams J W J 1964 Algorithms 232: Heapsort. *Commun. ACM* 7: 347–348

Wyman F B 1976 Improved event scanning mechanisms for discrete event simulation. *Commun. ACM* 19: 350–353

Zeigler B P 1976 *Theory of modelling and simulation* (New York: Wiley) (Reissued by Krieger, Malabar, FL in 1985)

Zeigler B P 1987 Hierarchical, modular, discrete-event modelling in an object-oriented environment. *Simulation* 49: 219–230

rallel algorithms for generating combinatorial objects on ear processor arrays with reconfigurable bus systems[*]

P THANGAVEL

Department of Computer Science, University of Madras, Chepauk,
Madras 600 005, India
e-mail: thang@unimad.ernet.in

Abstract. A bus system whose configuration can be dynamically changed is called reconfigurable bus system. In this paper, parallel algorithms for generating combinations, subsets, and binary trees on linear processor array with reconfigurable bus systems (PARBS) are presented.

Keywords. Parallel algorithms; reconfigurable bus systems; combinations; subsets; binary trees.

Introduction

nerating combinations and permutations are common combinatorial problems. We de-
e Cmn to be the number of combinations obtained by choosing m objects out of n
ects. Sequential algorithms optimally take a time of $O(m^*Cmn)$. Akl *et al* (1989/90)
e presented a parallel algorithm, using an Exclusive Read, Exclusive Write, Parallel
ndom Access Machine (EREW PRAM) model, which takes $O(Cmn)$ time optimally.
tion 3 considers the combinations generation using an m-processor linear PARBS. This
orithm outputs a combination in constant time. The combinations are generated in lex-
graphic ascending order. All the Cmn combinations can be generated and output in a
e of $O(Cmn)$. Though our algorithm takes the same time as that of Akl *et al* (1989/90),
present it because of the implementation feasibility of the PARBS model.
Generation of subsets has applications in subset sum, knapsack, base enumeration and
nimal covering problems (Stojmenovic & Miyakawa 1983). A cost-optimal systolic
orithm for generating subsets was presented in Tsay & Lee (1994). In § 4 we present a
t-optimal parallel algorithm using linear PARBS.
The problem of enumerating n-noded binary trees have been studied by introducing
eger sequences or permutations, which characterize trees in a combinatorial way. One

such sequence called P-sequence was introduced by Pallo & Racca (1985). They have provided a simple algorithm to generate P-sequences lexicographically. In § 5 an efficient and simple parallel algorithm for generating P-sequences is proposed which runs on a linear PARBS.

Zerling (1985) reported a new approach for generating rooted ordered binary trees with n-nodes in some order by using rotations. He established a 1–1 correspondence between binary trees and codewords related to rotations, called rotational admissible codewords for binary trees, and presented a recursive algorithm to generate these codewords. Er (1989) later presented faster recursive algorithm to generate these codewords. Recently Makinen (1991) has given a non-recursive algorithm. All the above algorithms take $O(n)$ time in the worst case to produce the next codeword from the previous one. In § 6 a simple parallel algorithm for generating rotational-admissible codewords based on the non-recursive algorithm of Makinen (1991) is proposed. This algorithm runs on linear PARBS.

In the next section the PARBS model is introduced.

2. The PARBS model

A reconfigurable bus system is a bus system whose configuration can be dynamically changed. Processors in PARBS are assumed to operate in single instruction multiple data (SIMD) mode; that is they perform the same instruction synchronously (at times only very few processors may be active but this will be explicitly specified in the instruction). In a single unit of time, each processor can either select a switch configuration, perform a read or write operation on the bus, or perform a constant-time computation on its local data. The model allows several processors to read the same bus component, but does not allow more than one processor to write on the same bus component at the same time. By giving specific instructions, the processors reading from a bus can be restricted. For example, only when a particular condition is satisfied the processor may read the bus. Each processor may have its own built-in ports. These ports can be dynamically connected in pairs based on computational needs. When a bus configuration is established, processors that are attached to the same bus can communicate with one another with the above restrictions. It is assumed that such a communication takes one unit of time. This basic assumption on PARBS models is acceptable because of the gap between signal propagation velocity and reconfiguring rates. With the current hardware technology, signal propagation velocity is very close to its limit (velocity of light), whereas the reconfiguration rates are bounded by both theoretical and technological restrictions. Thus the time for the switch state to change can be orders of magnitude longer than the time for the signal to traverse a distance equal to the diameter of a switch. It is also generally assumed that each processor can perform basic arithmetic and logical operations in unit time (Schuster 1991; Wang 1991).

The linear PARBS model can be described as follows: It consists of n-processors, P_1, $P_2, \ldots, P_N$. Each processor is connected to its neighbouring processors. The connections are made through two built-in ports denoted by N and S (or L and R) of each processor. A local switchable network could dynamically change the configuration of the system. It is controlled by the processor itself. Figure 1 shows such a system with $n = 5$.
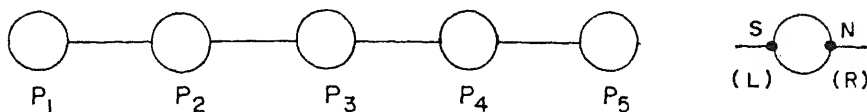
Figure 1. A 5-processor linear PARBS. Here P_i's – processors and S, N (L, R) are built-in ports.

3. Generation of combinations in a linear PARBS

Consider a linear PARBS with m-processors, each processor having six registers C, L, B, R, I and J. In the registers C and L, the components of current (which is used to generate the next one) and last combinations can be stored. Register B is used to store a boolean value by comparing C and L. Register R contains a copy of B of next processor. Registers I and J are used to store the indices of processors. Each combination is generated using its previous combination C_i, $i = 0, 1, 2, \ldots, m - 1$, and last combination L_i ($L_i = n - m + i + 1$), $i = 0, 1, 2, \ldots, m - 1$, in lexicographic order. Here, assume that we want to generate combinations of m integers from the first n integers ($1 < m < n$). Then it could be easily extended to an array of n elements by considering them as indices. The algorithm consists of the following steps.

Algorithm 1. Parallel generation of combinations.

Step 0: Initially $C_j = j + 1$; $L_j = n - m + j + 1$; $I_j = j$ and $J_j = 0$ are stored on corresponding registers of each processor P_j, $j = 0, 1, 2, \ldots, m - 1$.

Step 1: Set the boolean variable b_j with 1, if $C_j = L_j$, otherwise with 0 (if $C_j < L_j$) and store it in register B of each processor P_j, $0 \le j \le m - 1$.

Step 2: If $b_{m-1} = 0$, then set $C_{m-1} = C_{m-1} + 1$ and store it in register C of processor P_{m-1}. Go to step 8.

Step 3: If $b_0 = 1$ then, go to step 8 (the input combination is the last one i.e., same as L_j).

Step 4: Each processor P_j, $j = 1, 2, \ldots, m - 1$, sends a copy of b_j to processor P_{j-1}; say r_{j-1}, stored in R.

Step 5: Each processor P_i for which $b_i = 1$ connects port L to port R, for $1 < i < m - 1$.

Step 6: Processor P_j, for which $\bar{b}_j \cdot r_j = 1$, broadcasts C_j and I_j to processors P_k, $k = j+1$, $j + 2, \ldots, m - 1$, and received in registers C and J respectively.

Step 7: Each processor P_k for which $r_k = 1$ sets $C_k = C_k + I_k - J_k + 1$, and stored in register C.

Step 8: Each processor P_k, $k = 0, 1, \ldots, m - 1$ prints its register C value. Then go to step 1. □

To prove the correctness we proceed as follows: We can make the following observations from the lexicographic ascending order of combinations. If $C_0, C_1, \ldots C_{m-1}$ and

Proof. Let j be the smallest index such that

$$C_j = L_j \text{ (i.e., } b_1 = b_2 = \cdots = b_{j-1} = 0).$$

If $C_{j+1} \neq L_{j+1}$ then $C_{j+1} < L_{j+1}$ (by observation (ii)). From observation (i), $C_j < C_{j+1}$ and since $L_j < L_{j+1}$, we have $C_{j+1} \preceq L_j$ (L_j and L_{j+1} are consecutive), i.e., $C_{j+1} \preceq C_j$, which is a contradiction, therefore

$$C_{j+1} = L_{j+1}.$$

By induction $C_k = L_k$ for all $k > j$. □

If $b_0 = 1$ then $C_0 = L_0$, then $C_1 = L_1$, $C_2 = L_2, \ldots, C_{m-1} = L_{m-1}$ (by lemma 1). This implies that the current combination is the same as the last combination (step 3). If $b_{m-1} = 0$, there is no $C_j = L_j$, for $j = 0, 1, \ldots, m - 1$, then the processor P_{m-1} increments its component to produce the next combination in order (step 2). By lemma 1 there exists a unique index j such that the sequence $b_0, b_1, \ldots, b_{m-1}$ will be of the form $0, 0, \ldots, 0, 1, 1, \ldots, 1$, (first 1 is in processor P_{j+1}). This ensures that there is a unique processor P_j for which $\bar{b}_j \cdot r_j = 1$, therefore this processor can broadcast C_j and I_j in step 6. It also proves that step 5 can create a bus starting from processor P_j until processor P_{m-1}. To prove the correctness of step 7, we prove the following lemma.

Lemma 2. *Let $C_0, C_1, \ldots, C_{m-1}$ and $C1_0, C1_1, \ldots, C1_{m-1}$ be the current input and output combinations of the above algorithm. Then the two combinations are in lexicographic ascending order.*

Proof. Consider the boolean sequence

$$b_0, b_1, \ldots, b_{m-1} = 0, 0, \ldots, 0, 1, \ldots, 1,$$

(first 1 is at processor P_{j+1}). Then from the algorithm we can see that

$$C1_i = C_i \quad \text{for } i = 0, 1, \ldots, j - 1$$
$$C1_j = C_j + 1 > C_j$$

Hence the two combinations are in lexicographic ascending order. □

It can be easily seen that each step of the algorithm takes only a constant amount of time as the broadcast operation takes constant time. Thus we can generate the next combination from the previous one in constant time. The algorithm can be made adaptive (i.e., to run on an array with $k > m$ processors), if we divide the linear array into k/m groups of m processors such that each group can produce an interval of consecutive m-combinations. Note that in such cases we need to change the input (current and last) combinations correspondingly.

4. Parallel generation of subsets

We assume that each processor in the linear array has registers A, B, C, and D. Register A is used to store the elements of the current subset. B, C and D are used to store boolean

We shall present a parallel algorithm for generating all subsets in lexicographic order of the set $\{1, 2, \ldots, n\}$.

- Parallel generation of subsets.

$a_1 = 1, a_i = 0$, for $i > 1$, and $b_i = c_i = d_i = 0$, for all i, and store them in B, C and D respectively.

processor P_i prints a_i for which $a_i \neq 0$.

$= n$ then stop.

b_i with 1, if $a_i = 0$, and 0 otherwise, and c_i with 1 if $a_i = n$, and 0 otherwise.

processor P_i sends a copy of c_i to processor P_{i-1}, received in D, for $i = $ Processor P_i for which $c_i \cdot \bar{d_i} = 1$, sets $a_i = a_i + 1$.

cessor P_i for which $c_i = 1$, sets $a_i = 0$.

processor P_i for which $\bar{d_i} = 0$, connects port L with R. Processor P_i for 1 sends d_i value which is received in register D of P_1.

$D_1 = 0$, then each processor P_i sends a copy of register B value to P_{i+1} $< n$, received in register D. Then processor P_i for which $b_i \cdot \bar{d_i} = 1$ sets $+ 1$. Then go to step 2.   □

ch subset either contains only one n or none, there could be only one i for which is ensures that there could be atmost one processor for which $d_i = 1$ in step 7. e the entry prior to n is incremented by 1 (step 5) and the entry corresponding o zero (step 6) to get the next subset. Otherwise (if n is not a member of current e processor for which $b_i \cdot \bar{d_i} = 1$ sets its component of the next subset to $a_{i-1} + 1$. thm generates subsets of the set $\{1, 2, 3, 4\}$ in the following order:

			2			
2			2	3		
2	3		2	3	4	
2	3	4	2	4		
2	4		3			
3			3	4		
3	4		4			
4						

easily seen that the algorithm takes only constant time to generate the next subset previous one.

erating binary trees using P-sequences

shall introduce P-sequences. Let r_T be the degree of the root of the binary tree e by T_L and T_R, the left and right subtrees of T. Two orders, namely A-order and

DEFINITION 1

Two trees T and T' are in A-order, denoted by $T < T'$ if (1) $|T| < |T'|$ or (2) $|T| = |T'|$ and $T_L < T'_L$ or (3) $|T| = |T'|$ and $T_L = T'_L$ and $T_R < T'_R$ where $|T|$ is the number of leaves in T.

DEFINITION 2

Two trees T and T' are in B-order, denoted by $T < T'$ if (1) $r_T < r_{T'}$ or (2) $r_T = r_{T'}$ and $T_L < T'_L$ or (3) $r_T = r_{T'}$ and $T_L = T'_L$ and $T_R < T'_R$. P-sequences are used to generate binary trees in B-order. Now we define P-sequences as follows:

DEFINITION 3

For a given binary tree T, the P-sequence of T is the integer sequence $(p_T(1), p_T(2), \ldots, p_T(|T|-1))$ where $p_T(i)$ is the number of internal nodes O written before the leaf i in the Polish notation of T.

Theorem 1 (Pallo & Racca 1985). *An integer sequence $(p_1, p_2, \ldots, p_n)$ is the P-sequence of a binary tree with n internal nodes if and only if*

(1) *for all $i = 1, 2, \ldots, n$, $p_i < p_{i+1}$,*

(2) $p_n = n$, *and*

(3) *for all $i = 1, 2, \ldots, n$, $p_i > i$.*

Theorem 2 (Pallo & Racca 1985). *Given two binary trees T and T' such that $|T| = |T'|$, T and T' are in B-order if and only if p_T is lexicographically less than $p_{T'}$.*

The algorithm proposed here is based on the sequential generating algorithm in Pallo & Racca (1985). It is worth reproducing the same here.

Generating algorithm
    Begin with $(p_1, p_2, \ldots, p_n) = (1, 2, \ldots, n)$
        while $i = \max(k | p_k < n)$ exists do
            $p_i = p_{i+1}$
            for $j = i + 1$ to n do $p_j = \max(p_i, j)$ end do
        end do

We consider a linear PARBS with n processors. Each processor has registers B, C, D and I. B and C are used to store boolean values. D is used to store P-sequence elements and I is for index i of the corresponding processor P_i. Initially assume that $d_i = i$ and $I_i = i$ for $i = 1, 2, \ldots, n$. A parallel algorithm for generating P-sequences in lexicographic order is given in algorithm 3.

Algorithm 3. Parallel generation of P-sequences.

Step 1. Set $d_i = i$ and $I_i = i$ and store them in registers D and I of each processor P_i,

Step 3. If $b_1 = 1$ then stop.

Step 4. Each processor P_i sends a copy of b_i to processor P_{i-1}, say c_{i-1}, stored in C, for $i = 2, 3, \ldots, n$.

Step 5. Each processor P_i, for which $b_i = 1$, connects port L to port R.

Step 6. Processor P_i, for which $\bar{b}_i \cdot c_i = 1$, sets $d_i = d_{i+1}$ and broadcasts d_i to processors $P_{i+1}, P_{i+2}, \ldots, P_n$, stored in register D.

Step 7. Each processor P_i for which $b_{i=1}$, sets d_i with d_i, if $d_i > l_i$, and I_i otherwise.

Step 8. Each processor P_i writes d_i, $i = 1, 2, \ldots, n$ (next P-sequence). Then go to step 2.

□

Note that there is a unique processor P_i for which $\bar{b}_i \cdot c_{i=1}$, as the sequence b_i, $i = 1, 2, \ldots, n$ will be of the form $0, 0, \ldots, 0, 1, 1, \ldots, 1$ which follows from the definition of P-sequence. It takes only constant time to produce the next P-sequence from the previous one.

6. Parallel generation of binary trees using rotational admissible codewords

A rotational-admissible codeword is a sequence of integers obtainable by a rotation-based coding method. A codeword $(x_0, x_1, \ldots, x_{n-1})$ is rotational admissible if $x_0 = 0$ and for each $i = 1, 2, \ldots, n-1$, we have $0 \le x_i \le x_{i-1} + 1$. It has been shown that there is a one to one correspondence between codewords and binary trees (Zerling 1985). In this section we shall present a parallel algorithm based on Makinen's (1991) non-recursive algorithm, which is worth reproducing. It uses an array x containing the codewords to be generated, and x is initialised by zeroes. Basically the algorithm adds one to the previous codeword. Variable pos shows the position of x in which we perform the addition, and it is initialized to contain the value $n - 1$.

```
Procedure non-recursive Enumerate
   begin
      while x(0) = 0 do begin
         output codeword;
         if x(pos) ≤ x(pos-1) then
            x(pos) = x(pos) + 1
         else begin x(pos) = 0; pos = pos - 1;
            while x(pos) = x(pos-1) + 1 do
               begin x(pos) = 0; pos = pos-1 end;
            x(pos) = x(pos) + 1; pos = n - 1; end
      end
   end;
```

Consider a linear PARBS with each processor having registers A, B, C and X. X is used to store the components of the codeword. Register A is also used to store codewords. B and C are used to store boolean values. The algorithm is given below.

Algorithm 4. Parallel generation of rotational–admissible codewords

Step 1. Each processor P_i sets $x_i = 0$, for $0 \le i \le n - 1$.

Step 2. If $x_0 = 0$, then each processor P_i writes x_i, for $0 \leq i \leq n - 1$. Otherwise stop.

Step 3. Processor P_{n-2} sends a copy of x_{n-2} to processor P_{n-1}, received in register A.

Step 4. If $x_{n-1} \leq a_{n-1}$, processor P_{n-1} sets $x_{n-1} = x_{n-1} + 1$, and then go to step 2.

Step 5. Each processor P_i, $0 \leq i \leq n - 2$, sends a copy of x_i to processor P_{i+1}, received in register A.

Step 6. Processor P_{n-1} sets $x_{n-1} = 0$.

Step 7. Each processor P_i sets $a_i = a_i + 1$, stores in register A itself.

Step 8. Each processor P_i sets b_i with 1, if $a_i = x_i$ otherwise 0.

Step 9. Each processor P_i connects port L to port R for which $b_i = 1$. Processor P_{n-1} sends a 1 signal through this bus, which is received in register C of each processor in the bus. Disconnect all switches.

Step 10. Each processor P_i for which $\bar{b}_i \cdot c_i = 1$ sets $x_i = 0$ and $\bar{b}_i \cdot c_i = 1$ sets $x_i = x_i + 1$. Then go to step 2. □

It is worth noting that there is a unique processor P_i for which $\bar{b}_i \cdot c_i = 1$. The above algorithm is repeated until $x_0 = 0$. It can be easily seen that the algorithm takes only constant amount of time to produce the next codeword from the previous one.

7. Conclusion

We have presented parallel algorithms for generating combinatorial objects combinations, subsets, and binary trees based on P-sequences and rotational admissible codewords. Our algorithms take only constant time to produce next object from the previous one. There are many combinatorial open problems for which we can design fast parallel algorithms by exploiting the flexible nature of PARBS models.

References

Akl S G, Gries D, Stojmenovic I 1989/90 An optimal parallel algorithm for generating combinations. *Inf. Process. Lett.* 33: 135–139

Er M C 1989 A new algorithm for generationg binary trees using rotations. *Comput. J.* 32: 470–473

Makinen E 1991 Efficient generation of rotational-admissible codewords for binary trees. *Comput. J.* 34: 379

Pallo J, Racca R 1985 A note on generating binary trees in A-order and B-order. *Int. J. Comput. Math.* 18: 27–39

Schuster A 1991 *Dynamic reconfiguring networks for parallel computers: algorithms and complexity bounds.* PhD thesis, Hebrew University, Jerusalem

Stojmenovic I, Miyakawa M 1983 Applications of subset generating algorithm to base enumeration, knapsack and minimal covering problems. *Comput. J.* 1: 65–70

Tsay J C, Lee W P 1994 A cost-optimal systolic algorithm for generating subsets. *Int. J. Comput. Math.* 50: 1–10

Wang B F 1991 *Configurational computation: a new algorithm design strategy on processor arrays with reconfigurable bus systems.* PhD thesis, National Taiwan University, Taipei

Zerling D 1985 Generating binary trees using rotations. *J. Assoc. Comput. Mach.* 32: 694–701

Two cracks with coalesced interior plastic zones – The generalised Dugdale model approach

R R BHARGAVA and S C AGRAWAL

Department of Mathematics, University of Roorkee, Roorkee 247 667, India
e-mail: maths@rurkiu.ernet.in

Abstract. The problem investigated is of an elastic-perfectly plastic infinite plate containing two equal collinear and symmetrically situated straight cracks. The plate is subjected to loads at infinity inducing mode I type deformations at the rims of the cracks. Consequently, plastic zones are formed ahead of the tips of the cracks. The loads at infinity are increased to a limit such that the plastic zones formed at the neighbouring interior tips of the cracks get coalesced. The plastic zones developed at the tips of the cracks are closed by applying normal cohesive quadratically varying stress distribution over their rims. The opening of the cracks is consequently arrested. Complex variable technique is used in conjugation with Dugdale's hypothesis to obtain analytical solutions. Closed form analytical expressions are derived for calculating plastic zone size and crack opening displacement. An illustrative numerical example is discussed to study the qualitative behaviour of the loads required to arrest the cracks from opening with respect to parameters viz. crack length, plastic zone length and inter-crack distance. Crack opening displacement at the tip of the crack is also studied against these parameters.

Keywords. Plastic zone; mode I type deformation; opening mode deformation; rims of cracks; crack-opening displacement.

1. Introduction

Work on crack problems was first started by Inglis (1913). The complex variable formulation was given by Kolosov (1935). This was further developed by Muskhelishvili (1953) and was successfully applied by him to solve the crack problems for cracks of different geometry. The treatment of crack problems for fracture of brittle materials was given by Cherepanov (1974) using complex variable technique. Dugdale (1960) proposed an elastic–plastic model, called "strip yield model" for a sheet containing a slit. The sheet

opening was arrested by applying closing normal yield point stress to the rims of plastic zones. Smith (1974) gave a general theory based on the cohesive zone model for structures in the vicinity of a crack tip. The model was modified by Harrop (1978) for cases when plastic zones were closed by a cohesive normal parabolic stress distribution. The case of two interacting equal collinear straight cracks in an isotropic elastic unbounded plate was discussed by Viola (1983). Theocaris (1989) extended to "strip yield model" to the case when an infinitely elastic, perfectly plastic plate is weakened by two collinear straight cracks.

The present paper investigates the case of two equal, collinear, symmetrically-situated straight cracks weakening an infinite plate. The plate is subjected to uniform constant tension normal to the rims of the cracks, at the infinite boundary of the plate. Consequently, the cracks open forming plastic zones ahead of the tips of the cracks. The tension applied is increased to a limit such that the plastic zones at the two neighbouring interior tips of the cracks coalesce. The rims of plastic zones formed (at the interior and exterior tips of the cracks) are then subjected to normal cohesive stress distribution $t^2\sigma_{ye}$, where t is any point on any of the plastic zones and σ_{ye} is yield point stress. Thus, opening of the cracks is arrested.

2. Basic formulae

Stress components P_{xx}, P_{yy} and P_{xy} and displacement components u_x and u_y may be expressed in terms of two complex potentials $\phi(z)$ and $\Omega(z)$, developed by Muskhelishvili (1953), as

$$P_{yy} - i P_{xy} = \phi(z) + \Omega(\bar{z}) - (z - \bar{z})\phi'(z), \tag{1}$$

$$2\mu(u_{x,x} + i u_{y,x}) = \kappa\phi(z) - \Omega(\bar{z}) - (z - \bar{z})\phi'(z). \tag{2}$$

A bar over a function denotes its complex conjugate. A dash after a function denotes differentiation with respect to the argument, while a comma after a function signifies partial differentiation with respect to the subscript following it. The elastic constant μ denotes shear modulus and $\kappa = (3 - 4\nu)$ for the plane strain case and $\kappa = (1 - \nu)/(1 + \nu)$ for the plane stress case, ν being Poisson's ratio.

Consider a homogeneous, isotropic elastic infinite plate in the xy plane containing n straight cracks L_i ($i = 1, 2, \ldots n$) lying on the real x-axis. Dual problems of linear relationship are obtained using (1), when the rims of the cracks are acted upon by stresses $P_{yy}^{\pm}$ and $P_{xy}^{\pm}$. The two Hilbert problems so obtained may be written as

$$\phi^+(t) + \Omega^-(t) = P_{yy}^+ - i P_{xy}^+, \quad \text{on } L \tag{3}$$

$$\phi^-(t) + \Omega^+(t) = P_{yy}^- - i P_{xy}^-, \tag{4}$$

where $L = \sum_{i=1}^{n} L_i$, under the assumption $\lim_{y\to 0}\{y\phi'(t + iy)\} = 0$.

Superscripts $+$ and $-$ indicate the limiting value of the function when any point t on the crack, other than end points, is approached from the positive y-plane ($y > 0$) and the negative y-plane ($y < 0$), respectively.

The solutions of (3) and (4) for complex potentials $\phi(z)$ and $\Omega(z)$ may be written, as

$$\phi(z) = \phi_0(z) + (P_n(z)/X(z)), \tag{5}$$

$$\Omega(z) = \Omega_0(z) + (P_n(z)/X(z)), \tag{6}$$

where

$$\phi_0(z) = \int_L \frac{p(t)X(t)}{t-z}dt + \int_L \frac{q(t)X(t)}{t-z}dt - \frac{1}{2}\sigma_\infty, \tag{7}$$

$$\Omega_0(z) = \int_L \frac{p(t)X(t)}{t-z}dt - \int_L \frac{q(t)X(t)}{t-z}dt + \frac{1}{2}\sigma_\infty, \tag{8}$$

σ_∞ being the tension applied at infinite boundary,

$$p(t) = \frac{1}{2}[P_{yy}^+ + P_{yy}^-] - \frac{i}{2}[P_{xy}^+ + P_{xy}^-], \tag{9}$$

$$q(t) = \frac{1}{2}[P_{yy}^+ - P_{yy}^-] - \frac{i}{2}[P_{xy}^+ - P_{xy}^-], \tag{10}$$

and

$$X(z) = \prod_{i=1}^n [(z - a_i)(z - b_i)]^{1/2}, \tag{11}$$

a_i and b_i are the end points of the crack L_i and

$$P_n(z) = C_0 z^n + C_1 z^{n-1} + \cdots\cdots + C_n. \tag{12}$$

The constants C_i ($i = 1, 2, \ldots n$) are determined by the condition of single-valuedness of displacement at the rims of the crack and C_0 is determined from the boundary condition at infinity. Stress intensity factor, K_I ($= K_1 - iK_2$), at the crack tip $z = z_1$ may be calculated, as given by Cherepanov (1974), from

$$K_I = K_1 - iK_2 = 2\sqrt{2\pi} \lim_{z \to z_1} \{(z - z_1)^{1/2}\phi(z)\}, \tag{13}$$

where $\phi(z)$ is obtained from (5).

3. Formulation of the problem

A homogeneous, isotropic, elastic-perfectly plastic infinite plate, in the xy-plane, is weakened by two equal, collinear and symmetrically situated straight cracks L_1 and L_2. These cracks L_1 and L_2 occupy the ligaments $[-b, -a]$ and $[a, b]$ on the real x-axis as depicted in figure 1. The plate is subjected to uniform uniaxial tension, σ_∞, parallel to the y-axis, at infinite boundary. Consequently, the cracks L_1 and L_2 open forming the plastic zones at the tips $-b, -a, a$ and b. The prescribed tension at infinity is increased to the limit where the plastic zones developed at the tips $-a$ and a get coalesced. Thus the entire ligament $[-a, a]$ forms the interior plastic zone, Γ_1. The plastic zone Γ_2 ahead of the crack tip $-b$ occupies the ligament $[-c, -b]$ and the zone Γ_3 at the tip b occupies the ligament $[b, c]$. Each of the plastic zones Γ_i ($i = 1, 2, 3$) is closed by cohesive normal stress distribution $t^2\sigma_{ye}$, consequently arresting the opening of the cracks. Yield point stress is denoted by σ_{ye} and t is any point on any of the plastic zones.

4. Solution of the problem
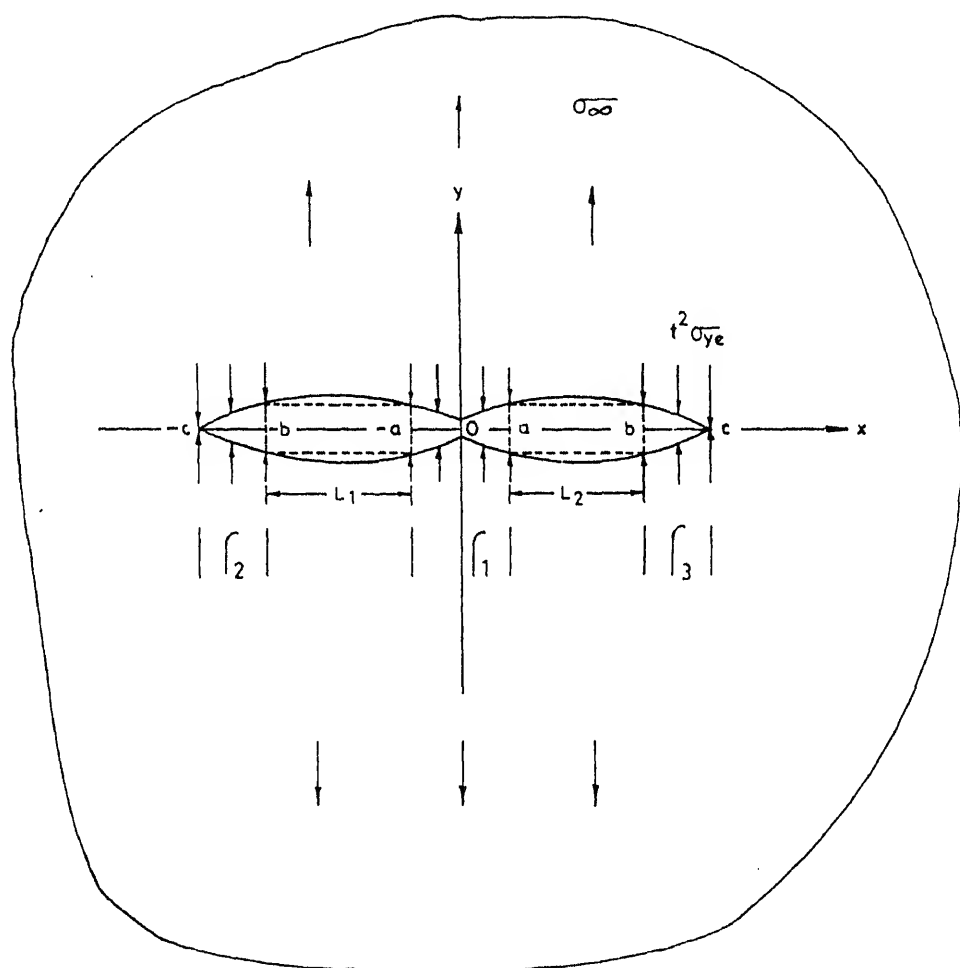
Figure 1. Configuration.

These two problems are appropriately derived from the problem stated in § 3. These are termed as *problem I* and *problem II*.

4.1 *Problem I*

The configuration of the problem I is: an infinite, homogeneous, isotropic and elastic-perfectly plastic infinite plate, in the xy-plane, contains a stress-free straight crack R_1 ($= \Gamma_2 U L_1 U \Gamma_1 U L_2 U \Gamma_3$) occupying ligament $[-c, c]$ of the real x-axis. Uniform tension σ_∞, parallel to y-direction, is applied at infinite boundary of the plate. The complex potentials $\phi_1(z)$, $\Omega_1(z)$ for this case may directly be written using Muskhelishvili (1953), as

$$\phi_1(z) = 0.5\sigma_\infty z(z^2 - c^2)^{-1/2} \qquad (14)$$
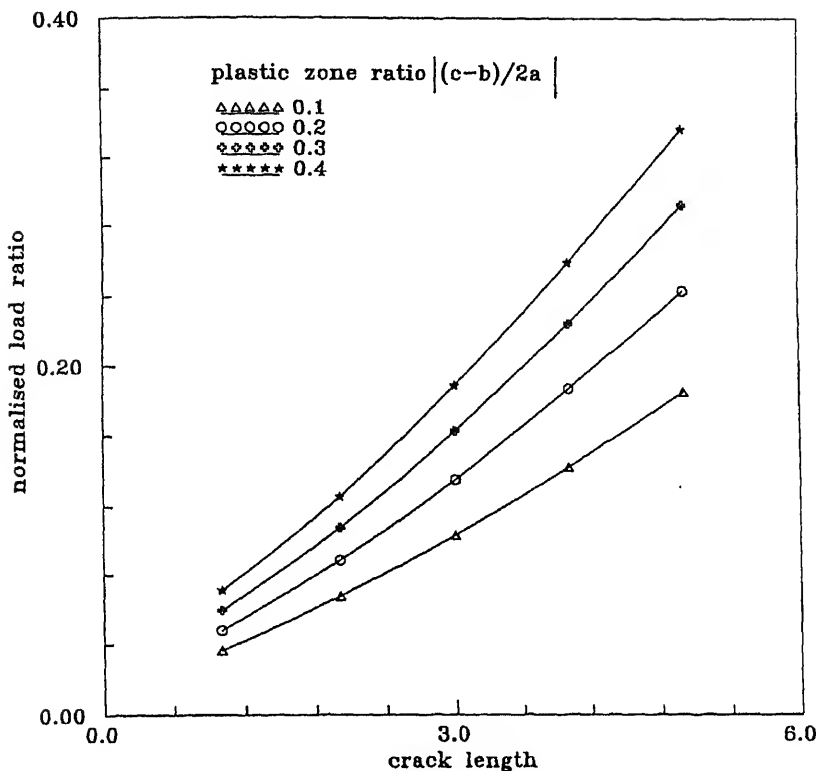
Figure 2. Variation of normalised load ratio versus crack length.

Opening mode stress intensity factor, K_1^I, for this problem, at the crack tip $z = c$ is calculated by substituting $\phi_1(z)$ for $\phi(z)$ from (14) into (13) and is given by Cherepanov (1974).

$$K_1^I = \sigma_\infty (\pi c)^{1/2} \tag{15}$$

4.2 *Problem II*

A homogeneous, isotropic and elastic-perfectly plastic infinite plate in the xy-plane contains two equal collinear symmetrically situated straight cracks L_1: $[-b, -a]$ and L_2: $[a, b]$ lying on the x-axis (refer figure 1). The cracks L_1 and L_2 possess the plastic zones Γ_1, Γ_2, Γ_3 ahead of the tips of the cracks. These plastic zones Γ_1, Γ_2, Γ_3 occupy the region $[-a, a]$, $[-c, -b]$ and $[b, c]$ respectively. The boundary conditions of the problem are the following.

(i) The unbounded plate is stress-free at infinite boundary.

(ii) The rims of the plastic zones Γ_i, $(i = 1, 2, 3)$ are subjected to tensile stress $P_{yy}^\pm =$
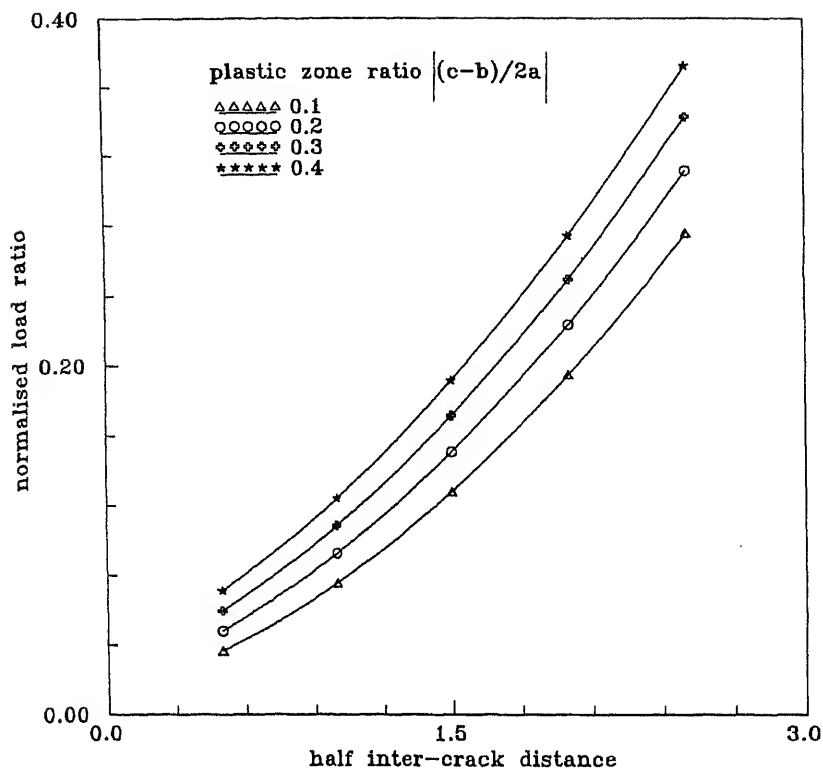
Figure 3.   Normalised load ratio versus inter-crack distance. Crack length is one unit.

Dual Hilbert problems are obtained in terms of potentials $\phi_2(z)$ and $\Omega_2(z)$ using (1) and boundary conditions (ii) and (iii) as

$$[\phi_2(t) + \Omega_2(t)]^+ + [\phi_2(t) + \Omega_2(t)]^- = t^2\sigma_{ye}, \tag{16}$$

$$[\phi_2(t) - \Omega_2(t)]^+ - [\phi_2(t) - \Omega_2(t)]^- = 0, \quad \text{on } \Gamma = U_{i=1}^3 \Gamma_i. \tag{17}$$

The subscript 2 indicates that the functions refer to the problem II.

Solutions of (16) and (17), are written using (5) to (12) and boundary conditions (i) and (iv), as

$$\phi_2(z) = \frac{\sigma_{ye}z}{\pi i X(z)} \left[\left(\frac{2c^2}{3} - z^2 \right) A_1 - izX(z)A_2 \right] \tag{18}$$

$$= \Omega_2(z),$$

$$A_1 = (\pi/2) + \sin^{-1}(a/c) - \sin^{-1}(b/c),$$

$$A_2 = \tan^{-1} \frac{z(c^2 - a^2)^{1/2}}{a(z^2 - c^2)^{1/2}} - \tan^{-1} \frac{z(c^2 - b^2)^{1/2}}{b(z^2 - c^2)^{1/2}} - \frac{\pi}{2},$$

where,

$$X(z) = (c^2 - z^2)^{1/2}. \tag{19}$$
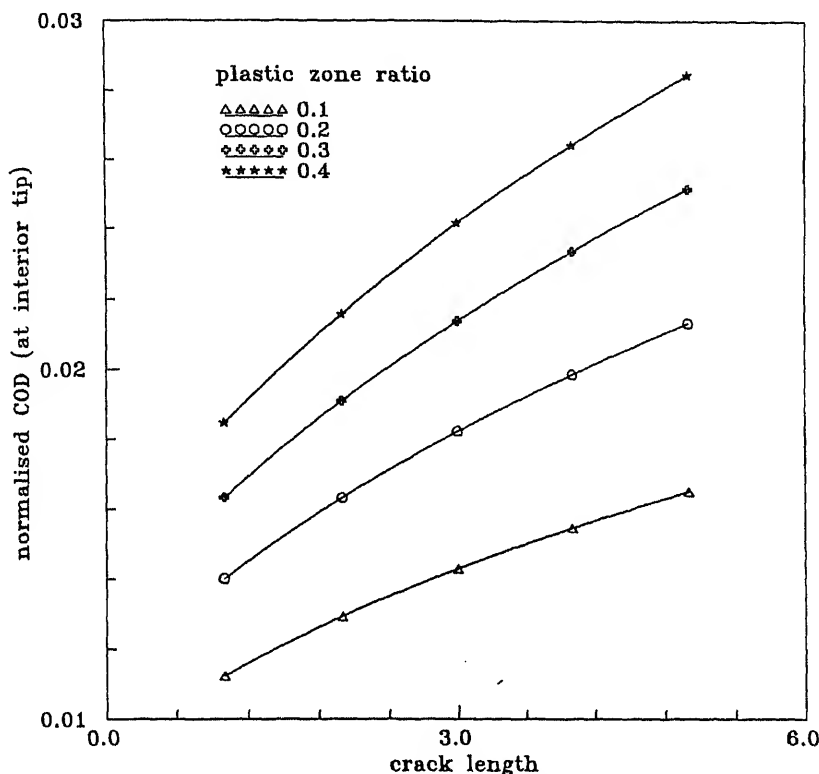
Figure 4. Variation of normalised crack opening displacement versus crack length at the interior tip of the crack.

Opening mode stress intensity factor for problem II, K_1^{II}, is obtained substituting the lue of $\phi_2(z)$ for $\phi(z)$ from (18) into (13) and simplifying, one gets

$$K_1^{II} = -\frac{2\sigma_{ye}c^2}{3}\left(\frac{c}{\pi}\right)^{1/2}. \tag{20}$$

Plastic zone size and crack opening displacement

astic zone size for original problem (stated in §3) is obtained by superposing two stress tensity factors obtained for problem I and problem II. Using the fact that the stresses main finite at every point of the plate leads to the nonlinear equation.

$$\frac{c^2}{3}[1 - (2/\pi)\{\sin^{-1}(b/c) - \sin^{-1}(a/c)\}] - \sigma_\infty/\sigma_{ye} = 0, \tag{21}$$

r determining c when values of $\sigma_\infty/\sigma_{ye}$, a and b are prescribed. Finally the plastic zone ngth is calculated by $|c - b|$.

The Dugdale model crack face opening displacement. $2u$ corresponding to the calcu-
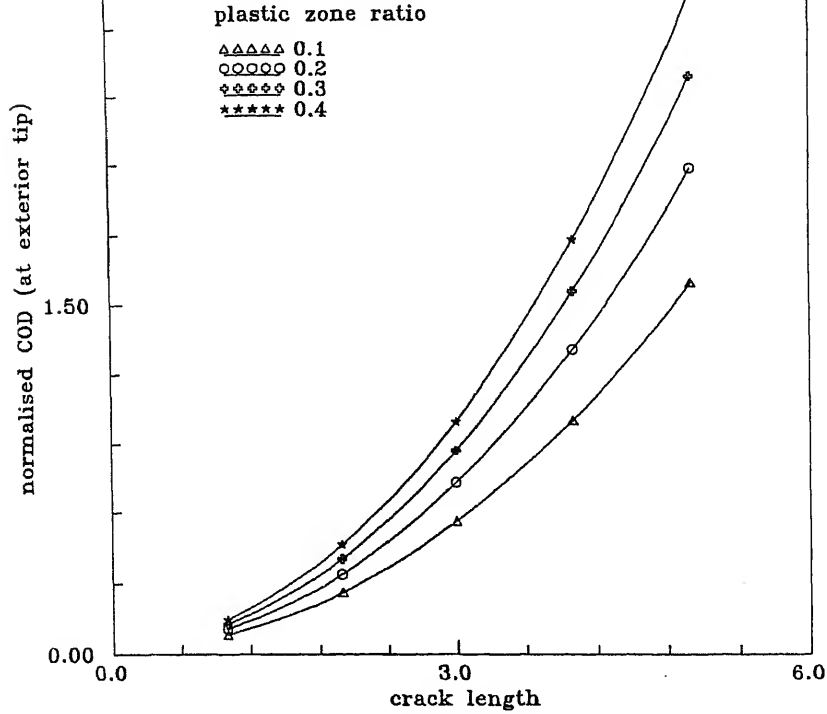
Figure 5. Variation of normalised crack opening displacement versus crack length at the exterior tip of the crack.

where E is Young's modulus, Im[ ] denotes the imaginary part of the quantity in the bracket. The complex potential $\Phi(z)$ is obtained by superposing the non-singular terms of $\phi_1(z)$ {from (14)} and $\phi_2(z)$ {from (18)}. Substituting the value of $\Phi(z)$ thus obtained in (22) and integrating, one obtains the crack-opening displacement (COD) at the exterior tip $z = b$ as

$$u_y = -(2\sigma_{ye}cb/\pi E)A_1\{b(1 - b^2/c^2)^{1/2} + c\sin^{-1}(b/c)\}, \tag{23}$$

and at the interior tip $z = a$ the COD is given by

$$u_y = -(2\sigma_{ye}ca/\pi E)A_1\{a(1 - a^2/c^2)^{1/2} + c\sin^{-1}(a/c)\}. \tag{24}$$

6. Illustrative example

A qualitative study has been carried out to find the load ratio required to prevent the cracks formed from opening, as the crack length and inter-crack distance (the distance between the two neighbouring tips of the two cracks) is increased.

Normalised load ratio has been plotted against increasing crack length in figure 2. Normalising factor of stress intensity factor (SIF) is 10. It is observed that for longer cracks greater load is required as expected to close the crack. Also for a fixed crack length,
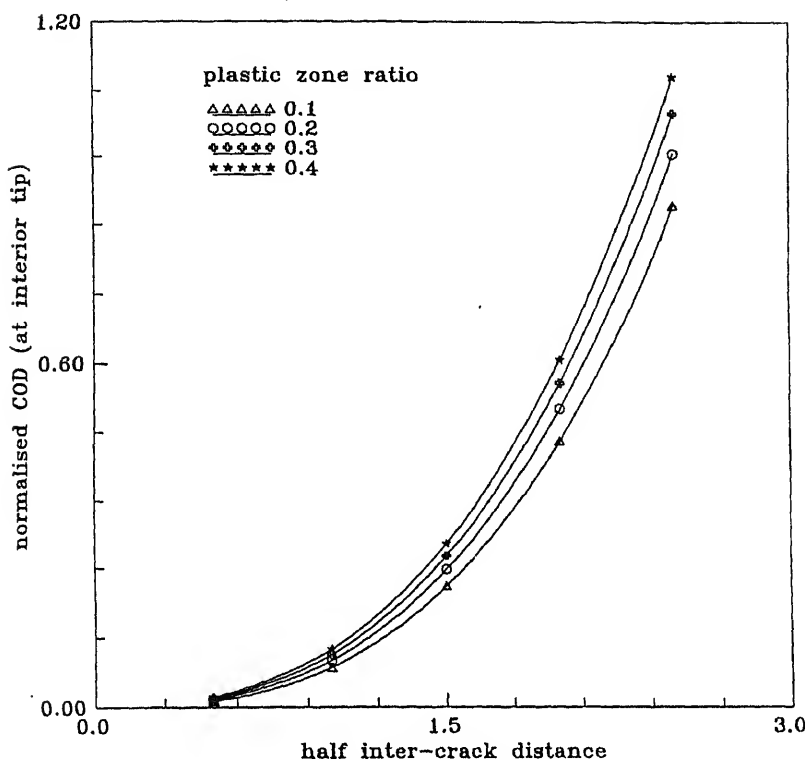
Figure 6. Normalised crack opening displacement at interior tip with respect to inter-crack distance. The crack length is taken as one unit.

if plastic zone ratio (the ratio of outer plastic zones/inner plastic zone) is increased then a higher load still is required to close the crack. The two cracks were kept apart at a fixed distance.

Figure 3 depicts the variation of load ratio as inter-crack distance is increased. The crack length is unity. It may be noted that, for a fixed plastic zone ratio, as the cracks move apart greater load is required to close them. This is because the effect of the cracks on one another diminishes.

The variation of crack-opening displacement at the interior tip, $z = a$ of the crack L_2, versus increase in crack length is depicted in figure 4. The normalising factor of COD is $(\sigma_{ye}/10E)$. The same variation is plotted at the exterior tip $z = b$ of the crack L_2 in figure 5. It may be noted as crack length increases the cracks open more at both the tips, for a fixed plastic zone ratio. But it is interesting to note the difference in the variation patterns of COD at the inner and outer tips of the crack. It may be noted that at the two interior tips of the COD, the pattern is such that the plastic zones get coalesced.
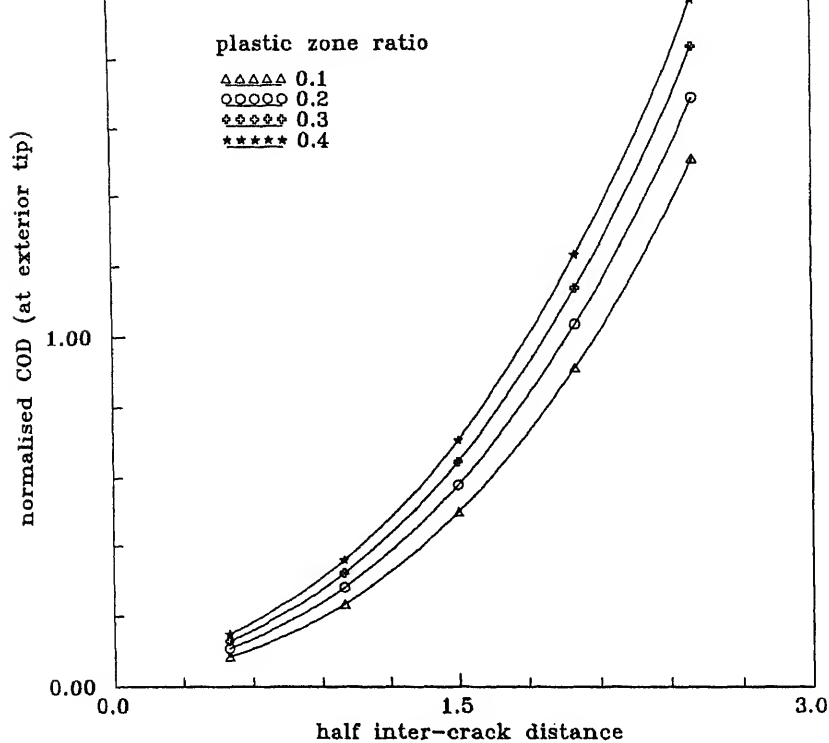
Figure 7. Normalised crack opening displacement at exterior tip with respect to inter-crack distance. The crack length is one unit.

7. Conclusion

The Dugdale model (1960) was modified by Harrop (1978) for a single crack when a parabolic load is distributed over its plastic zones. The present paper modifies the extended Dugdale model given by Theocaris (1989) for two straight cracks in an elastic-perfectly plastic infinite plate. Here, the quadratically varying cohesive stress distribution closes the rims of the plastic zone.

The results obtained may be applied to a plate-type structure in which two similar straight cracks develop in the vicinity of each other and their propagation under pressure causes coalescence in the plastic zone, viz. aeroplane wings, turbine blades, stretched sheets etc.

The dimensions of the plate should be about five times larger than the total length of the two cracks and the inter-crack distance should be such that the stresses at the crack tips do not affect the stresses at the boundary of the plate.

List of symbols

a, b, c	tips of the cracks;
E	Young's modulus;
K_I	stress intensity factor for mode I type deformations;
L_j $(j = 1, 2)$	cracks;
P_{ij} $(i, j = x, y)$	stress components;
$P_n(z), X(z)$	complex functions;
u_i $(i = x, y)$	displacement components;
$z = x + iy$	complex variable;
Γ_j $(j = 1, 2, 3)$	plastic zones;
μ	shear modulus;
ν	Poisson's ratio;
σ_{ye}	yield point stress;
σ_∞	tension applied at infinite boundary;
$\phi(z), \Omega(z), \Phi(z)$	complex potentials.

References

Cherepanov G P 1974 *Mechanics of brittle fracture* (First English Transl.) (New York: McGraw Hill)

Dugdale D S 1960 Yielding steel sheets containing slits. *J. Mech. Phys. Solids.* 8: 100–104

Harrop L P 1978 Application of a modified Dugdale model for two K vs COD relation. *Eng. Fracture Mech.* 10: 807–816

Inglis C E 1913 Stresses in a plastic due to presence of cracks and sharp corners. *Trans. Inst. Naval Architects* 55: 219–241

Kolosov G V 1935 *Application of the complex variable to the theory of elasticity* (Moscow Leningrad: ONTI)

Muskhelishvili N I 1953 *Some basic problems of the mathematical theory of elasticity* (ed.) J R M Radok (Groningen: Noordhoff)

Smith E 1974 The structure in the vicinity of a crack tip. A general theory based on the cohesive zone model. *Eng. Fracture Mech.* 6: 213–222

Theocaris P S 1989 Dugdale model for two collinear unequal cracks. *Eng. Fracture Mech.* 18: 213–222

Viola E 1983 Non singular stress effects on two interacting equal collinear cracks. *Eng. Fracture Mech.* 18: 801–814

Analysis of deformed microstrip resonator using the finite element method

A S CHAUDHARI and P B PATIL[*]

Department of Physics, Dr B A M University, Aurangabad 431 004, India

Abstract. Microstrip resonator with shape deformation has been analysed using the finite element method (FEM). Keeping the front surface of the microstrip resonator fixed, the back surface is deformed. Five cases of the deformation of the back surface are considered: (i) the left vertical side of the back surface is shifted inward; (ii) both vertical sides of the back surface are shifted inward; (iii) the right vertical side of the back surface is shifted outward; (iv) both vertical sides of the back surface are shifted outward; (v) left vertical side of back surface is shifted inward and the right vertical side of it is shifted outward. Variation in cutoff frequency for the TM_{011}, TM_{110}, TM_{111}, TM_{012}, TM_{112} and TM_{210} is observed.

Keywords. Microstrip resonator; finite element method.

1. Introduction

Finite element method (FEM) is a powerful tool for the approximate solutions of differential equations governing diverse physical problems. It can tackle problems of different branches in physics and it can be used for any complicated geometry or domain of regular or irregular shape. The method can deal with the problems described by coupled equation. The method is useful for inhomogeneous and anisotropic media also.

The resonant frequencies of open microstrip ring resonators are determined by Wolff & Menzel (1975) and Pintzos & Pregla (1978). The microstrip resonator with equilateral triangular patch is studied by Wolf & Knoppic (1974), Helszajn *et al* (1979), Lyon & Helszajn (1982). The triangular and rectangular patch microstrip resonator has been analysed by Kalamse & Patil (1994) using the finite element method.

In this paper we have analysed the microstrip resonator with shape deformation for five cases:

(i) the left vertical side of the back surface is shifted inward;
(ii) both vertical sides of the back surface are shifted inward;

---
*For correspondence

(iii) the right vertical side of the back surface is shifted outward;

(iv) both vertical sides of the back surface are shifted outward;

 (v) left vertical side of back surface is shifted inward and the right vertical side of it is shifted outward.

2. Statement of the problem

Consider the rectangular microstrip resonator bounded by six faces B1, B2, B3, B4, B5, B6, as shown in figure 1a.

The two side surfaces B_1, B_2 and front and back surfaces B3, B4 are magnetic walls. The top and the bottom surfaces are electric walls.

The electric field **E** within the resonator will satisfy Maxwell's equations,

$$\text{Curl Curl } \mathbf{E} - K^2 \mathbf{E} = 0, \tag{1}$$

where $K^2 = \omega^2 \mu_0 \varepsilon_0$

$$\text{grad div } \mathbf{E} - \nabla^2 \mathbf{E} - K^2 \mathbf{E} = 0. \tag{2}$$

Since the medium is charge free, div **E** = 0.

$$\nabla^2 \mathbf{E} + K^2 \mathbf{E} = 0. \tag{3}$$

The electric field within the substrate has only the Z component and the magnetic field has the X and Y components. The tangential component of the magnetic field at the edge is negligible,
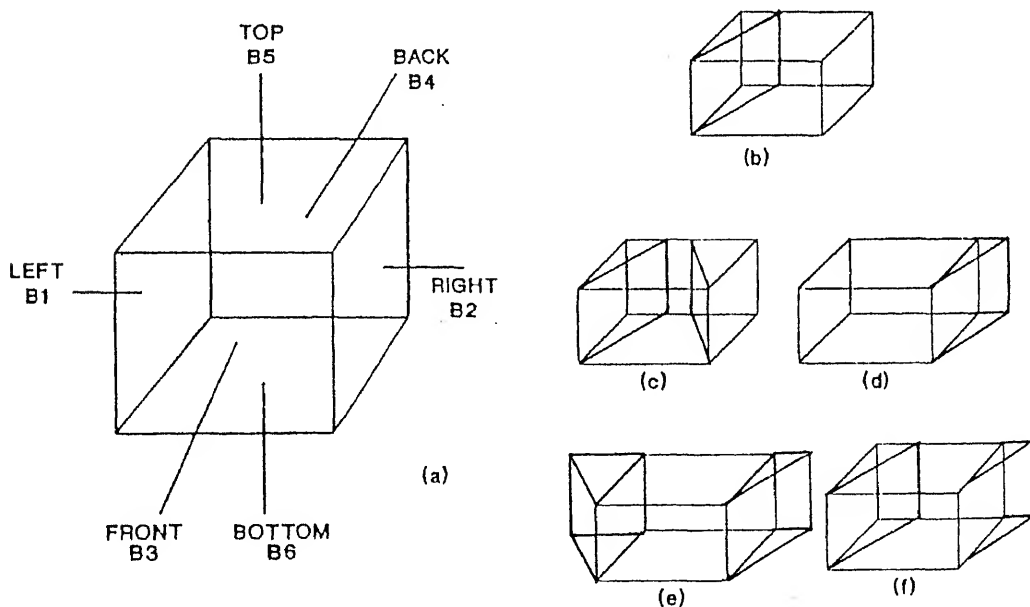
Figure 1. (a) Microstrip resonator without deformation. (b) Left vertical side of the back surface is shifted inward. (c) Both vertical sides of the back surface are shifted inward. (d) Right vertical side of the back surface is shifted outward. (e) Both vertical sides of the back surface are shifted outward. (f) Left vertical side of the back surface is shifted inward and right vertical side of the back surface is shifted outward.

$$\text{div grad } E_z + K^2 E_z = 0. \tag{4}$$

The fields within the resonator corresponding to TM modes will be generated by the
equation,

$$\nabla^2 E_z + K^2 E_z = 0, \tag{5}$$

subjected to the boundary condition

$$\left.\frac{\partial E_z}{\partial n}\right|_{B_1,B_2,B_3,B_4} = 0, \tag{6}$$

$$E_z|_{B_5,B_6} = 0, \tag{7}$$

where E_z is the z component of E, and $\partial/\partial n$ represents normal derivative.

Variational formulation

To get the expression for the functional Π in variational formulation, multiply (4) by some
weight function V and integrate it over the domain of the resonator,

$$\Pi = \int\int_\Omega\int V \text{ div grad } E_z \, d\Omega + K^2 \int\int_\Omega\int V E_z \, d\Omega. \tag{8}$$

Using the vector identity $S \text{ div } \mathbf{A} = \text{div}(S\mathbf{A}) - (\text{grad } S) \cdot \mathbf{A}$ for the first term, (8) becomes

$$\Pi = \int\int_\Omega\int \text{div}(V \text{ grad } E_z) \, d\Omega - \int\int_\Omega\int (\text{grad } V)(\text{grad } E_z) \, d\Omega$$
$$+ K^2 \int\int_\Omega\int V E_z \, d\Omega. \tag{9}$$

Applying Gauss Divergence theorem to the first term in (9), we get

$$\Pi = - \int\int_\Omega\int (\text{grad } V) \cdot (\text{grad } E_z) \, d\Omega + K^2 \int\int_\Omega\int V E_z \, d\Omega$$
$$+ \int_S\int V \text{ grad } E_z \cdot \bar{n} \, ds. \tag{10}$$

Using (6), the last term in (10) will vanish.

Table 1. Variation of cutoff frequency for different modes for the inner deformation of one vertical side.

Shift	Frequency (GHz)					
(cm)	TM$_{011}$	TM$_{110}$	TM$_{111}$	TM$_{012}$	TM$_{112}$	TM$_{210}$
05	13.607	14.016	15.005	16.470	17.641	17.906
0	13.606	14.046	15.035	16.470	17.663	17.992
5	13.605	14.077	15.066	16.470	17.685	18.073
20	13.604	14.107	15.100	16.470	17.707	18.148
25	13.603	14.139	15.136	16.470	17.728	18.219
30	13.601	14.170	15.175	16.470	17.749	18.287
35	13.599	14.202	15.217	16.471	17.770	18.351
40	13.597	14.233	15.261	16.473	17.792	18.413

Table 2. Variation of cutoff frequency for different modes for the inner deformation of both vertical sides.

Shift	Frequency (GHz)					
L (cm)	TM_{011}	TM_{110}	TM_{111}	TM_{012}	TM_{112}	TM_{210}
0.05	13.607	14.045	15.035	16.469	17.667	17.990
0.10	13.607	14.103	15.104	16.468	17.719	18.130
0.15	13.607	14.161	15.184	16.466	17.778	18.269
0.20	13.608	14.218	15.277	16.464	17.842	18.379
0.25	13.608	14.272	15.386	16.461	17.915	18.477
0.30	13.609	14.324	15.513	16.458	17.998	18.567
0.35	13.611	14.372	15.659	16.455	18.093	18.650
0.40	13.612	14.417	15.825	16.451	18.203	18.728

By substituting $V = E_z^*$ and changing the sign, (10) becomes

$$\Pi = \frac{1}{2} \int \int_\Omega \int (\text{grad } E_z) \cdot (\text{grad } E_z^*) \, d\Omega - K^2 \frac{1}{2} \int \int_\Omega \int E_z E_z^* \, d\Omega. \qquad (11)$$

$\frac{1}{2}$ is introduced since Π is bilinear functional.

The first variation $\delta \Pi$ is given by,

$$\delta \Pi = \delta \frac{1}{2} \left[\int \int_\Omega \int (\nabla E_z) \cdot (\nabla E_z^*) \, d\Omega - K^2 \int \int_\Omega \int E_z E_z^* \, d\Omega \right]. \qquad (12)$$

For Π to be stationary, $\delta \Pi$ should be minimum.

4. Discretization

According to the finite element method (Reddy 1986; Akin 1988), the volume of the resonator is divided into hexahedral elements with 20 nodes. The mapping functions assumed for these elements are quadratic in nature. The functional over an element is given by

$$\Pi^e = \sum_{\text{ele}} \frac{1}{2} \{E_z^e\}^T [S^e] \{E_z^e\} - \frac{1}{2} K^2 \sum_{\text{ele}} \{E_z^e\}^T [T^e] \{E_z^e\}, \qquad (13)$$

Table 3. Variation of cutoff frequency for different modes for the outer deformation of one vertical side.

Shift	Frequency (GHz)					
L (cm)	TM_{011}	TM_{110}	TM_{111}	TM_{012}	TM_{112}	TM_{210}
0.05	13.607	13.957	14.950	16.470	17.595	17.719
0.10	13.606	13.930	14.925	16.469	17.570	17.622
0.15	13.604	13.903	14.902	16.469	17.551	17.515
0.20	13.602	13.877	14.879	16.469	17.522	17.416
0.25	13.599	13.853	14.857	16.468	17.495	17.315
0.30	13.595	13.831	14.836	16.467	17.468	17.214
0.35	13.590	13.810	14.816	16.465	17.438	17.116
0.40	13.583	13.792	14.797	16.463	17.410	17.019

ble 4. Variation of cutoff frequency for different modes for the outer deformation of both vertical les.

ift	Frequency (GHz)					
(cm)	TM_{011}	TM_{110}	TM_{111}	TM_{012}	TM_{112}	TM_{210}
05	13.607	13.928	14.926	16.469	17.574	17.620
10	13.607	13.872	14.882	16.468	17.533	17.416
15	13.607	13.817	14.843	16.465	17.494	17.211
20	13.607	13.765	14.808	16.461	17.457	17.014
25	13.607	13.714	14.777	16.453	17.422	16.826
30	13.606	13.666	14.749	16.437	17.388	16.656
35	13.606	13.621	14.723	16.388	17.355	16.528
40	13.606	13.577	14.700	16.269	17.323	16.479

here

$$S_{ij} = \int \left[\frac{\partial F_i}{\partial x} \frac{\partial F_j}{\partial x} + \frac{\partial F_i}{\partial y} \frac{\partial F_j}{\partial y} + \frac{\partial F_i}{\partial z} \frac{\partial F_j}{\partial z} \right] dx\, dy\, dz \qquad (14)$$

d

$$T_{ij} = \int F_i F_j \, dx\, dy\, dz. \qquad (15)$$

Here F_i is the mapping function due to the ith node and integrations are over the mesh ement surface.

The functional for the whole region Ω is given by

$$\Pi = \frac{1}{2} \{E_z\}^T [S]\{E_z\} - \frac{1}{2} K^2 \{E_z\}^T [T]\{E_z\}. \qquad (16)$$

he condition that variation of Π must be minimum i.e. zero, gives

$$[S]\{E_z\} - K^2 [T]\{E_z\} = 0. \qquad (17)$$

quation (17) is the matrix equation to be solved to get eigenvalues and eigenvectors.

Numerical calculations for microstrip resonator

microstrip resonator of sides $2.4\,\text{cm} \times 1.2\,\text{cm} \times 2.8\,\text{cm}$ is considered. The deformation r each case is shown in figure 1(b–f).

able 5. Variation of cutoff frequency for different modes for the inner deformation of one vertical side d outer deformation of other vertical side.

ift	Frequency (GHz)					
(cm)	TM_{011}	TM_{110}	TM_{111}	TM_{012}	TM_{112}	TM_{210}
05	13.606	13.987	14.976	16.470	17.615	17.816
10	13.603	13.991	14.973	16.471	17.605	17.820
15	13.599	13.996	14.969	16.472	17.588	17.827
20	13.593	14.004	14.964	16.474	17.565	17.836
25	13.586	14.013	14.956	16.476	17.537	17.848
30	13.577	14.025	14.948	16.479	17.505	17.862
35	13.568	14.037	14.938	16.482	17.469	17.879
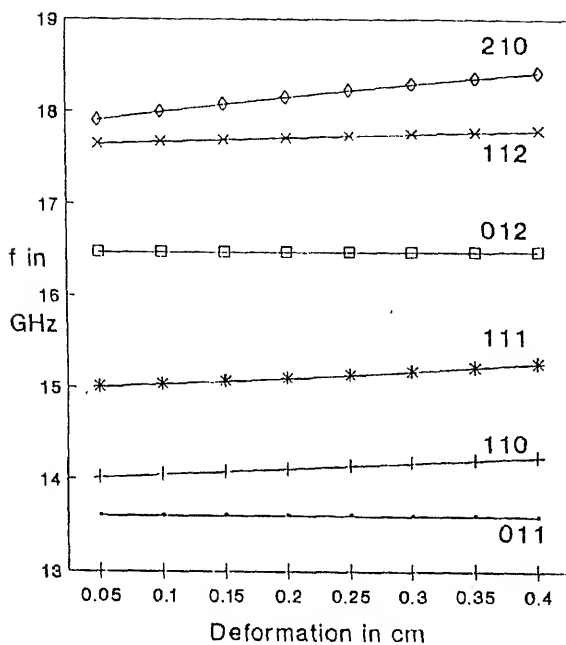
Figure 2. Variation of cutoff frequencies for different modes for the inner deformation of one vertical side of back surface.
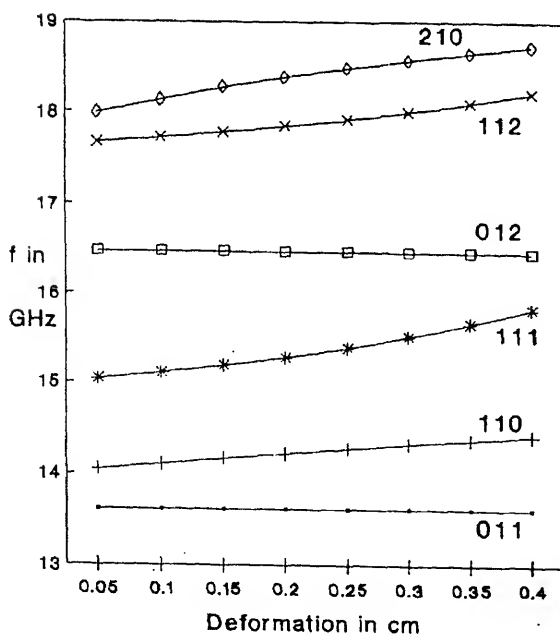
Figure 3. Variation of cutoff frequencies for different modes for the inner deformation of both vertical sides of back surface.
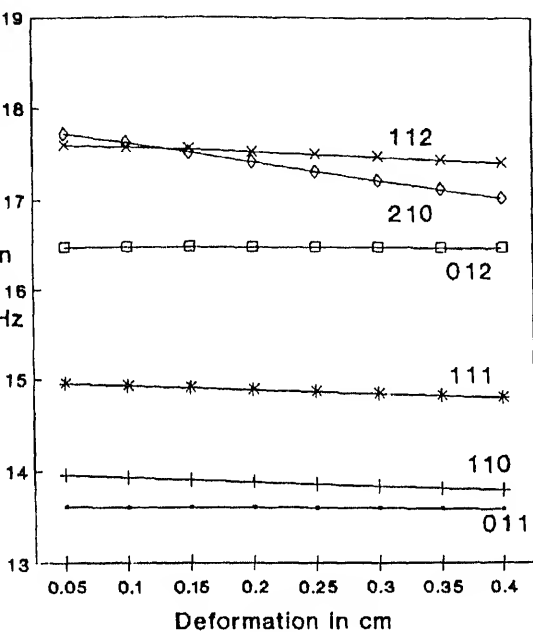
Figure 4. Variation of cutoff frequencies for different modes for the outer deformation of one vertical side of back surface.
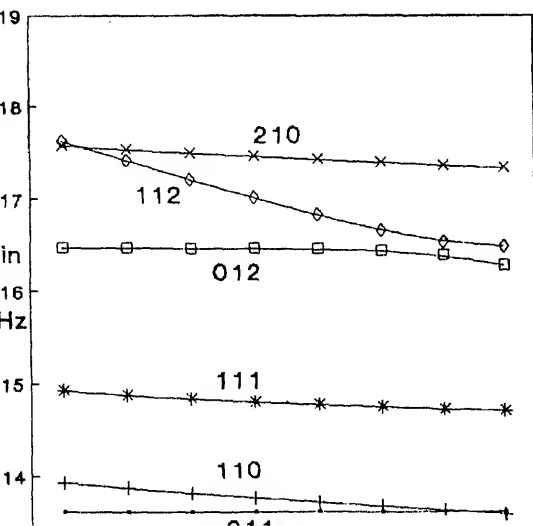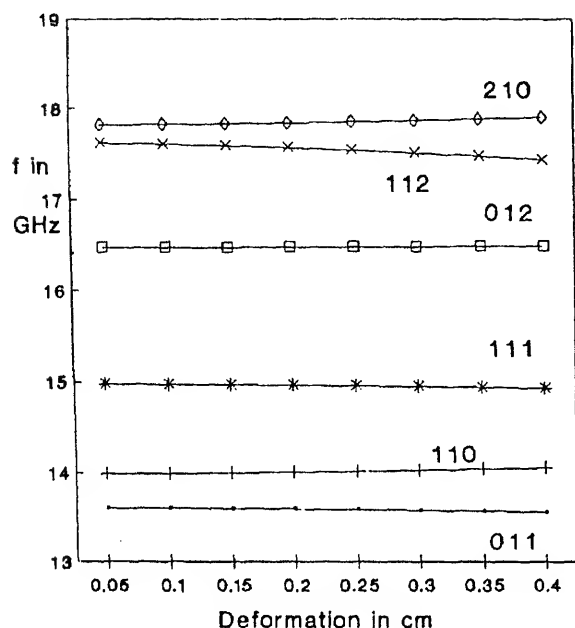
Figure 6. Variation of cutoff frequencies for different modes for the inner deformation of one vertical side and outer deformation of other vertical side of back surface.

(i) First case: Left vertical side of the back surface of the microstrip resonator is shifted inward by 0.05 cm.

(ii) Second case: Both vertical sides of the back surface of the microstrip resonator are shifted inward by 0.05 cm.

(iii) Third case: Right vertical side of the back surface of the microstrip resonator is shifted outward by 0.05 cm.

(iv) Fourth case: Both vertical sides of the back surface of the microstrip resonator are shifted outward by 0.05 cm.

(v) Fifth case: Left vertical side of the back surface is shifted inward by 0.05 cm and the right vertical side is shifted outward by 0.05 cm.

In each case, for every shift, eigenvalues and eigenvectors are obtained. TM_{011}, TM_{110}, TM_{111}, TM_{012}, TM_{112} and TM_{210} modes are identified using field plots. These eigenvalues are the values of $\omega^2 \mu_0 \varepsilon_0$. From this, the corresponding cutoff frequencies are calculated in gigahertz.

For every shift, in each case, variation of cutoff frequency with the shift for the above modes is given in tables 1–5 and shown graphically in figures 2–6.

6. Conclusions

For the inner deformation of one or both vertical sides of the back surface, the cutoff frequencies increase with the shift for all the modes, except for TM_{011} and TM_{012} where they are almost constant.

For the outer deformation of one or both vertical sides of the back surface, the cutoff quencies decrease with the shift for all the modes.
For the inward deformation of one vertical side and outward deformation of the other tical side, values of cutoff frequency for the TM_{011}, TM_{111} and TM_{112} modes decrease th the shift, while for the TM_{110}, TM_{012} and TM_{210} modes they increase.

ferences

in J E 1988 *Application and implementation of finite element method* (New York: Academic Press)

lszajn J, James D S, Nisbet W T 1979 Circulators using planar triangular resonators. *IEEE Trans. Microwave Theory Tech.* 27: 188–193

lmase G M, Patil P B 1994a Finite element analysis of triangular patch microstrip resonator with shape deformation. *Indian J. Phys.* B68: 341–345

lmase G M, Patil P B 1994b Finite element analysis of rectangular patch microstrip resonator with shear deformation. *Indian J. Pure Appl. Phys.* 32: 899–901

on R W, Helszajn J 1982 A finite element analysis of plane circulators using arbitrary shaped resonators. *IEEE Trans. Microwave Theory Tech.* 30: 1964–1974

tzos S G, Pregla R 1978 A simple method for computing the resonant frequencies of microstrip ring resonator. *IEEE Trans. Microwave Theory Tech.* 26: 809–813

ddy J N 1986 *An introduction to finite element method* (New York: McGraw Hill)

lff I, Knoppik N 1974 Rectangular and circular microstrip capacitors and resonator. *IEEE Trans. Microwave Theory Tech.* 22: 857–864

lff I, Menzel W 1975 The microstrip ring resonator. *IEEE Trans. Microwave Theory Tech.* 23: 441–444

Effect of surface stresses on surface waves in elastic solids

PRANABES KANTI PAL [1], D ACHARYA [2] and P R SENGUPTA [3]

[1] Indian Institute of Mechanics of Continua, 201, Manicktola Main Road, Flat no 42, Calcutta 700 054, India
[2] Department of Mathematics, Mahadevananda College, Burrackpore 743 101, India
[3] Department of Mathematics, University of Kalyani, Kalyani 741 235, India

Abstract. This paper deals with the propagation of surface waves in homogeneous, elastic solid media whose free surfaces or interfaces of separation are capable of supporting their own stress fields. The general theory for the propagation of surface waves in a medium which supports surface stresses is first deduced, and then this theory is employed to investigate the particular cases of surface waves, viz. (a) Rayleigh waves, (b) Love waves and (c) Stoneley waves. It is seen that the Rayleigh waves become dispersive in nature; and, in case of low frequency with residual surface tension, a critical wavelength exists, below which the propagation of Rayleigh waves is not possible. This critical wave length is directly proportional to the surface tension. Some numerical calculations have been made in the case of Love waves and conclusions have been drawn.

Keywords. Surface waves; effect of surface stresses; Rayleigh waves; Love waves; Stoneley waves; residual stress.

1. Introduction

Surface waves play an important role in the science of earthquakes. To match the actual situation in a given problem, three types of surface waves are generally introduced which are classified according to their nature of displacements (Dey & Sengupta 1978; Chandrasekharaiah 1986; Das *et al* 1992; Pal *et al* 1996). Again it is known that the physical properties of bodies in the neighbourhood of the surface are sensibly different from those in the interior. Thus the boundary surface may be regarded as a two-dimensional

propagation. The Rayleigh wave velocity equation has been obtained as a particular case. Following Chandrasekharaiah (1987), the effect of surface stresses on Rayleigh wave propagation has been discussed. Next we discuss Love wave propagation with surface stress on its free boundry, with some numerical computations and conclusions. Wave velocity equation for Stoneley waves has also been deduced.

2. Basic equations

Let us consider two homogeneous, isotropic, elastic solid media M_1 (lower medium) and M_2 (upper medium) with different material constants. We also consider that a two-dimensional elastic layer is present at the surface of separation of the two media. We take an orthogonal cartesian frame of axes $Ox_1x_2x_3$, O is taken on the interface and Ox_3 is positive in the downward direction. As in Bullen & Bolt (1985), here also the wave travels in the Ox_1 direction. The disturbances of the particles on lines parallel to Ox_2 vibrate in phase, which means all partial derivatives with respect to x_2 are zero.

The equation of motion in the absence of body forces is

$$\mu \nabla^2 \mathbf{u} + (\lambda + \mu)\, \text{grad div } \mathbf{u} = \rho \frac{\partial^2 \mathbf{u}}{\partial t^2}. \tag{1}$$

In the above Navier's equation of classical elasticity $\mathbf{u}\ (u_1, u_2, u_3)$ is the displacement vector; λ, μ are Lamé elastic constants; ρ is mass density and t represents time. To investigate plane deformation parallel to the x_1x_3 plane we introduce the displacement potentials P, Q related to the displacement components $u_1(x_1, x_3, t)$ and $u_3(x_1, x_3, t)$ by the equations,

$$u_1 = \frac{\partial P}{\partial x_1} - \frac{\partial Q}{\partial x_3}, \quad u_3 = \frac{\partial P}{\partial x_3} + \frac{\partial Q}{\partial x_1}. \tag{2}$$

Introducing (2) in (1) we obtain,

$$\left(\nabla^2 - \frac{1}{a^2} \frac{\partial^2}{\partial t^2} \right) P = 0, \tag{3}$$

$$\left(\nabla^2 - \frac{1}{b^2} \frac{\partial^2}{\partial t^2} \right) Q = 0, \tag{4}$$

$$\left(\nabla^2 - \frac{1}{b^2} \frac{\partial^2}{\partial t^2} \right) u_2 = 0, \tag{5}$$

where $a = (\lambda + 2\mu/\rho)^{1/2}$ and $b = (\mu/\rho)^{1/2}$ are velocities of dilatation and distortion respectively. Befitting the actual situation of the problem, we seek solutions to (3)–(5) as

$$[P, Q, u_2] = [\bar{P}(x_3), \bar{Q}(x_3), \bar{u}_2(x_3)]\, e^{i(\eta x_1 - \zeta t)}. \tag{6}$$

Inserting (6) in (3)–(5), we get

$$\frac{d^2 \bar{P}}{dx_3^2} - \eta^2 \left(1 - \frac{\zeta^2}{\eta^2 a^2}\right) \bar{P} = 0,$$

$$\frac{d^2 \bar{Q}}{dx_3^2} - \eta^2 \left(1 - \frac{\zeta^2}{\eta^2 b^2}\right) \bar{Q} = 0, \tag{7}$$

$$\frac{d^2 \bar{u}_2}{dx_3^2} - \eta^2 \left(1 - \frac{\zeta^2}{\eta^2 b^2}\right) \bar{u}_2 = 0.$$

Now, $\bar{P}(x_3)$, $\bar{Q}(x_3)$, $\bar{u}_2(x_3)$ describe surface waves and as such they must be vanishingly small as $x_3 \to \infty$. Hence the solution of (5) for the medium M_1 may be taken as

$$P = A_1 \exp(-\eta m_1 x_3)\, e^{i(\eta x_1 - \zeta t)},$$

$$Q = A_2 \exp(-\eta m_2 x_3)\, e^{i(\eta x_1 - \zeta t)}, \tag{8}$$

$$u_2 = A_3 \exp(-\eta m_2 x_3)\, e^{i(\eta x_1 - \zeta t)},$$

where

$$m_1 = (1 - r^2)^{1/2}, \quad m_2 = (1 - s^2)^{1/2},$$

$$c = \frac{\zeta}{\eta}, \quad r = \frac{c}{a}, \quad s = \frac{c}{b}, \tag{9}$$

wavelength $= 2\pi/\eta$, wave velocity $= \zeta/\eta$, ζ being a known constant while η is an unknown constant; and, A_1, A_2, A_3 are arbitrary constants.

Analogous solutions with primes (for the region $0 \le x_3 < -\infty$) are

$$P' = A_1' \exp(\eta m_1' x_3)\, e^{i(\eta x_1 - \zeta t)},$$

$$Q' = A_2' \exp(\eta m_2' x_3)\, e^{i(\eta x_1 - \zeta t)}, \tag{10}$$

$$u_2' = A_3' \exp(\eta m_2' x_3)\, e^{i(\eta x_1 - \zeta t)},$$

are obtained for the upper medium M_2 of the material, where

$$m_1' = (1 - r'^2)^{1/2}, \quad m_2' = (1 - s'^2)^{1/2},$$

$$c = \frac{\zeta}{\eta}, \quad r' = \frac{c}{a'}, \quad s' = \frac{c}{b'}. \tag{11}$$

3. Boundary conditions and solutions

We assume that the plane $x_3 = 0$ is a material layer which adheres, without slipping, to the lower medium and is a two-dimensional elastic continuum. This layer is capable of supporting its own stress represented by surface stress tensor $\Sigma_{i\alpha}$, and obeys the law (Chandrasekharaiah 1987),

$$\sum = \begin{cases} \delta_{i\alpha}[\sigma + (\lambda_0 + \sigma)\, u_{\gamma,\gamma}] + \mu_0 u_{i,\alpha} + (\mu_0 - \sigma) u_{\alpha,i}, & \text{for } i, \alpha, \gamma = 1, 2, \\ \sigma u_{3,\alpha}, & \text{for } i = 3. \end{cases}$$

Here λ_0, μ_0 are the Lamé moduli of the material boundary and σ is the residual surface tension on the layer $x_3 = 0$. Following Gurtin & Murdoch (1976) we use the surface elasticity tensor. The forces on the bounding surface are governed by surface stress tensor $\sum_{i\alpha}$. Given a smooth oriented curve γ in $x_3 = 0$ with positive unit normal n_i, $\sum_{i\alpha} n_\alpha$ is the force per unit length of γ by that part of the boundary into which n_α is directed, upon that part of $x_3 = 0$ away from which n_α points. Dimensions of λ_0, μ_0, σ are the same (N/m).

The boundary conditions are

(i) $$\tau_{i3} + \sum_{i\alpha,\alpha} -\rho_0 \frac{\partial^2 u_i}{\partial t^2} = \tau'_{i3}, \quad \text{on } x_3 = 0,$$

where ρ_0 is the mass per unit surface area (kg m^{-2}) of the layer (Gurtin & Murdoch 1976), and τ_{ij} and τ'_{ij} are the stress tensors in the interior of the media M_1 and M_2 respectively. The conventional stress tensor τ_{i3} has the unit of force per unit area and the new stress tensor $\sum_{i\alpha}$ has the unit of force per unit length. τ_{ij} obeys the law (Gurtin & Murdoch 1976),

$$\tau_{ij} = \lambda \delta_{ij} u_{k,k} + \mu(u_{i,j} + u_{j,i}). \tag{13}$$

The boundary conditions are

$$\mu \left(2\frac{\partial^2 P}{\partial x_1 \partial x_3} + \frac{\partial^2 Q}{\partial x_1^2} - \frac{\partial^2 Q}{\partial x_3^2} \right) + \left[(\lambda_0 + 2\mu_0)\frac{\partial^2}{\partial x_1^2} - \rho_0 \frac{\partial^2}{\partial t^2} \right] \left(\frac{\partial P}{\partial x_1} - \frac{\partial Q}{\partial x_3} \right)$$

$$= \mu' \left(2\frac{\partial^2 P'}{\partial x_1 \delta x_3} + \frac{\partial^2 Q'}{\partial x_1^2} - \frac{\partial^2 Q'}{\partial x_3^2} \right), \tag{14}$$

$$2\mu \left(\frac{\partial^2 Q}{\partial x_1 \partial x_3} - \frac{\partial^2 P}{\partial x_1^2} \right) + \frac{\mu}{b^2}\frac{\partial^2 P}{\partial t^2} + \left(\alpha\frac{\partial^2}{\partial x_1^2} - \rho_0 \frac{\partial^2}{\partial t^2} \right) \left(\frac{\partial P}{\partial x_3} + \frac{\partial Q}{\partial x_1} \right)$$

$$= 2\mu' \left(\frac{\partial^2 Q'}{\partial x_1 \partial x_3} - \frac{\partial^2 P'}{\partial x_1^2} \right) + \frac{\mu'}{b'^2}\frac{\partial^2 P'}{\partial t^2} \left(\mu_0 \frac{\partial^2}{\partial x_1^2} - \rho_0 \frac{\partial^2}{\partial t^2} + \mu\frac{\partial}{\partial x_3^2} \right) u_2$$

$$= \mu' \frac{\partial^2 u'_2}{\partial x_3^2}.$$

(ii) The displacement components must be continuous on the surface of separation $x_3 = 0$ at all places and times. Therefore

$$\frac{\partial P}{\partial x_1} - \frac{\partial Q}{\partial x_3} = \frac{\partial P'}{\partial x_1} - \frac{\partial Q'}{\partial x_3},$$

$$\frac{\partial P}{\partial x_3} + \frac{\partial Q}{\partial x_1} = \frac{\partial P'}{\partial x_3} + \frac{\partial Q'}{\partial x_1}, \tag{14a}$$

$$u_2 = u'_2.$$

Applying boundary conditions (14) and (14a) to (8) and (10) the following system of equations is obtained,

$$(2m_1 + \eta F)A_1 - i(2 - s^2 + \eta m_2 F)A_2$$
$$+ \frac{2\mu'}{\mu}m_1'A_1' + i\frac{\mu'}{\mu}(2 - s'^2)A_2' = 0, \tag{15}$$

$$(2 - s^2 + \eta m_1 H)A_1 - i(2m_2 + \eta H)A_2$$
$$- \frac{\mu'}{\mu}(2 - s'^2)A_1' - i2\frac{\mu'}{\mu}m_2'A_2' = 0,$$

$$A_1 - im_2 A_2 - A_1' - im_2'A_2' = 0,$$

$$m_1 A_1 - iA_2 + m_1'A_1' + iA_2' = 0,$$

$$\left(\mu_0 - \rho_o c^2 + \frac{\mu}{\eta}m_2\right)A_3 + \frac{\mu'}{\eta}m_2'A_3' = 0,$$

$$A_3 - A_3' = 0,$$

where we have taken

$$\frac{1}{\mu}(\lambda_0 + 2\mu_0 - \rho_0 c^2) = F,$$
$$\frac{1}{\mu}(\sigma - \rho_0 c^2) = H. \tag{16}$$

The last two equations of (15) yield

$$A_3 = A_3' = 0,$$

which implies that there is no propagation of displacement u_2. Eliminating the constants A_1, A_2, A_1', A_2' from (15) we obtain

$$\Delta = \det[a_{ij}] = 0, \quad i, j = 1, 2, 3, 4, \tag{17}$$

where

$$a_{11} = 2m_1 + \eta F, \ a_{12} = 2 - s^2 + \eta m_2 F,$$

$$a_{13} = 2\frac{\mu'}{\mu}m_1', \ a_{14} = \frac{\mu'}{\mu}(2 - s'^2),$$

$$a_{21} = 2 - s^2 + \eta m_1 H, \ a_{22} = 2m_2 + \eta H,$$

$$a_{23} = -\frac{\mu'}{\mu}(2 - s'^2),$$

$$a_{24} = -2\frac{\mu'}{\mu}m_2',$$

$$a_{31} = 1, \ a_{32} = m_2, \ a_{33} = -1, \ a_{34} = -m_2',$$

$$a_{41} = m_1, \ a_{42} = 1, \ a_{43} = m_1', \ a_{44} = 1.$$

Equation (17) represents the wave velocity dispersion equation for interface waves in elastic solid media under the influence of surface stresses, where the plane of separation

$x_3 = 0$ is a material boundary. The equation depends on the particular value of η which indicates dispersive nature caused by the presence of surface stresses. If the surface of separation is free of surface stresses, $F = H = 0$; and at once the classical equation of general surface waves is obtained from (17).

4. Particular cases

4.1 *Rayleigh waves*

As a particular case of the general surface wave discussed in the previous article we may study Rayleigh waves under the influence of surface stresses in the light of the investigation done by Chandrasekharaiah (1987). Here the upper medium (M_2) is replaced by vacuum. Applying the boundary conditions

$$\tau_{13} + \sum_{1\alpha,\alpha} -\rho_0 \frac{\partial^2 u_1}{\partial t^2} = 0,$$

$$\tau_{33} + \sum_{3\alpha,\alpha} -\rho_0 \frac{\partial^2 u_3}{\partial t^2} = 0,$$

we get

$$(2m_1 + \eta F)A_1 + (2 - s^2 + \eta m_2 F)A_2 = 0,$$
$$(2 - s^2 + \eta m_1 H)A_1 + (2m_2 + \eta H)A_2 = 0.$$

Eliminating the constants A_1, A_2 the Rayleigh wave velocity equation may be obtained as

$$\begin{vmatrix} 2m_1 + \eta F & 2 - s^2 + \eta m_2 H \\ 2 - s^2 + \eta m_1 H & 2m_2 + \eta H \end{vmatrix} = 0.$$

The determinantal equation on simplification gives

$$(1 - m_1 m_2)FH\eta^2 + (m_1 H + m_2 F)s^2\eta + 4m_1 m_2 - (2 - s^2)^2 = 0. \tag{18}$$

This equation is in agreement with the corresponding equation obtained by Chandrasekharaiah (1987) in the absence of voids. Again in the case of conventional stress-free boundary, (18) becomes

$$4m_1 m_2 - (2 - s^2)^2 = 0. \tag{19}$$

When surface stresses $\sum_{i\alpha}$ arise from surface tension, we have as a special case the conditions $\lambda_0 = \mu_0 = \rho_0 = 0$. In such a situation $F = 0$, $H = \sigma/\mu$ and hence the equation (18) becomes

$$\eta\sigma s^2(1 - s^2\tau_0^2)^{1/2} = \mu[(2 - s^2)^2 - 4(1 - s^2)^{1/2}(1 - s^2\tau_0^2)^{1/2}], \tag{20}$$

where

$$\tau_0^2 = b^2/a^2. \tag{21}$$

This equation is in agreement with the corresponding equation obtained by Chandrasekharaiah (1987) in the absence of voids. In the absence of residual surface tension, we take $\sigma = 0$ to obtain the classical Rayleigh wave velocity equation (19). In this case, (20) has one root $s = s_0$ in the open interval $(0, 1)$. If $\sigma \neq 0$, (20) has a real root in the interval $(s_0, 1)$, provided $\eta < \eta_c$ where $\eta_c = \mu/[\sigma(1 - \tau_0^2)^{1/2}]$. Again if $\eta = \eta_c$ we see that $s = 1$ is a root of (20). Lastly, the said equation has no real root if $\eta > \eta_c$. Thus we have a critical wavelength, $L_c = 2\pi/\eta_c$, below which Rayleigh waves cannot propagate at all under the influence of surface stresses. The relation,

$$(\mu/2\pi)L_c = \sigma(1 - (b^2/a^2))^{1/2},$$

reveals that critical wavelength varies directly with residual surface tension which is a factor of surface stress.

4.2 Love waves

In order to consider Love waves we shall assume that the upper medium (M_2) is of finite thickness H and the lower medium (M_1) is semi-infinite. Let the origin O of the orthogonal cartesian frame $Ox_1x_2x_3$ be on the interface, Ox_1 and Ox_3 are along the interface and vertically downwards into M_1 respectively. For Love wave propagation along x_1 axis (see figure 1),

$$u_1 = u_3 = 0,$$
$$u_2 = u_2(x_1, x_3, t).$$

Here equations of motion are

$$\left(\nabla^2 - \frac{1}{b^2}\frac{\partial^2}{\partial t^2}\right)u_2 = 0,$$
$$\left(\nabla^2 - \frac{1}{b'^2}\frac{\partial^2}{\partial t^2}\right)u_2' = 0. \tag{22}$$
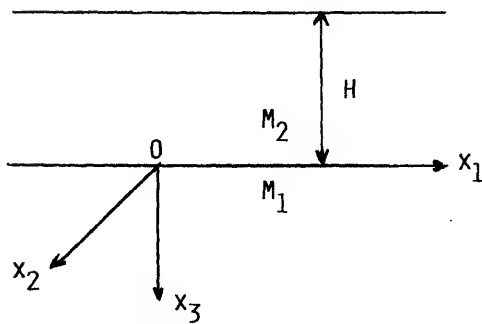
Figure 1. Love waves and their propagation.

$$u_2 = A_1 \exp(-\eta m_2 x_3)\, e^{i(\eta x_1 - \zeta t)},$$
$$u_2' = \{A_2 \exp(-\eta m_2' x_3) + A_3 \exp(\eta m_2' x_3)\}\, e^{i(\eta x_1 - \zeta t)}, \tag{23}$$

where

$$m_2 = (1 - s^2)^{1/2},$$
$$m_2' = (1 - s'^2)^{1/2}, \tag{24}$$
$$C = \frac{\zeta}{\eta}, \quad s = \frac{C}{b}, \quad s' = \frac{C}{b'}.$$

The boundary conditions are:

(i) u_2 is continuous on the boundary surface of separation $x_3 = 0$;

(ii) τ_{23} is continuous on the boundary surface of separation $x_3 = 0$;

(iii) since the boundary surface $x_3 = -H$ is free of external loads

we have (Gurtin & Murdoch 1974–75)

$$\tau_{23}' + \sum_{2\alpha,\alpha} = \rho_0 \frac{\partial^2 u_2'}{\partial t^2} \quad \text{for } x_3 = -H.$$

Applying the above boundary conditions to (23) and then eliminating the constants A_1, A_2, A_3 we get the wave velocity equation for Love waves under the influence of surface stresses in determinantal form as,

$$\begin{vmatrix} 1 & -1 & -1 \\ \mu m_2 & -\mu' m_2' & \mu' m_2' \\ 0 & \left\{\dfrac{\mu' m_2'}{\eta} + (\mu_0 - \rho_0 c^2)\right\} e^{\eta m_2' H} & -\left\{\dfrac{\mu' m_2'}{\eta} - (\mu_0 - \rho_0 c^2)\right\} e^{-\eta m_2' H} \end{vmatrix} = 0.$$

This determinantal equation on simplification gives

$$\eta H = \frac{1}{\{(c/b')^2 - 1\}^{1/2}}$$
$$\times \tan^{-1}\left[\frac{\{(c^2/b'^2) - 1\}^{1/2}\{(\mu/\mu')[1 - (c^2/b^2)]^{1/2} + (\eta/\mu')(\mu_0 - \rho_0 c^2)\}}{(c^2/b'^2) - 1 - (\mu/\mu')[1 - (c^2/b^2)]^{1/2} \cdot (\eta/\mu')(\mu_0 - \rho_0 c^2)}\right]. \tag{25}$$

when the surface stress effect is neglected ($\mu_0 = \rho_0 = 0$), (25) reduces to

$$\eta H = \frac{1}{\{(c^2/b'^2) - 1\}^{1/2}} \tan^{-1}\left[\frac{(\mu'/\mu)\{1 - c^2/b^2\}^{1/2}}{\{(c^2/b'^2) - 1\}^{1/2}}\right], \tag{26}$$

which is the classical Love wave velocity equation. From (25) we observe that Love wave propagation is possible if

$$1 < (c/b') < [(\mu/\rho)/(\mu'/\rho')]^{1/2}. \tag{27}$$

Table 1. Values of ηH.

η	1.048	1.096	1.144	1.192	1.240	1.288	1.336	1.384	1.432
1	6.933	4.081	2.858	2.147	1.671	1.325	1.057	0.838	0.648
2	7.444	4.576	3.333	2.602	2.104	1.733	1.439	1.193	0.973
3	7.623	4.757	3.517	2.788	2.291	1.922	1.628	1.381	1.159
0*	2.975	1.630	1.078	0.771	0.571	0.428	0.317	0.225	0.139

(column group header: c/b')

* Denotes the classical case where the surface stresses are not present

4.3 Numerical calculation of Love waves

For numerical calculations we take

$$\mu = 3.00 \times 10^6 \, \text{N/cm}^2, \quad \mu' = 5.00 \times 10^6 \, \text{N/cm}^2, \quad \mu_0 = 6.47 \times 10^6 \, \text{N/cm},$$
$$\rho = 2.72 \, \text{g/cm}^3, \quad \rho' = 9.89 \, \text{g/cm}^3, \quad \rho_0 = 3.40 \, \text{g/cm}^2.$$

Table 1 gives the values of ηH for different values of C/b' taking $\eta = 1, 2, 3$. It also gives the values of ηH for different values of C/b' in the absence of surface stresses (denoted by 0 in the first column).

We observe that the Love wave velocity equation is independent of σ and depends solely upon μ_0 and ρ_0. It is also observed that for a particular value of η, wave velocity decreases with increase of ηH. Again for a particular wave velocity, ηH increases with increase of η. For comparison we give the values of ηH in table 1 for the classical case (denoted by 0) where the surface stresses are not present.

4.4 Stoneley waves

Stoneley waves are a generalized form of Rayleigh waves propagating along the common boundary of M_1 and M_2. Equation (17) represents the wave velocity equation for Stoneley waves in elastic media with surface stresses. This equation after a little manipulation may be obtained in the form

$$\eta^2 CD(1 - ST)(1 - S'T')$$
$$+ s^2 \left\{ (CT + DS)(1 - S'T') + \frac{\mu'}{\mu}\tau_2^2(CT' + DS')(1 - ST) \right\} \eta - \Delta(s) = 0, \quad (28)$$

where

$$\Delta(s) = (1 - S'T') \left\{ (2 - s^2)^2 - 4ST - 4\frac{\mu'}{\mu}(1 - s^2 - ST) \right\}$$
$$+ (1 - ST) \left\{ (2 - \tau_2^2 s^2)^2 - 4S'T' - 4\frac{\mu'}{\mu} - (1 - \tau_2^2 s^2 - S'T') \right\}$$
$$- \tau_2^2 s^4 \frac{\mu'}{\mu}(2 + ST' + S'T), \quad (29)$$

and

$$s = c/b, \quad C = \sigma^*/\rho b^2, \quad \sigma^* = \sigma - \rho_0 c^2,$$

$$D = \frac{\Gamma^*}{\rho b^2}, \quad \Gamma^* = \Gamma - \rho_0 c^2, \quad \Gamma = \lambda_0 + 2\mu_0,$$

$$S = (1 - s^2)^{1/2}, \quad T = (1 - \tau_1^2 s^2)^{1/2},$$

$$S' = (1 - \tau_2^2 s^2)^{1/2}, \quad T' = (1 - \tau_3^2 s^2)^{1/2},$$

$$\tau_1 = b/a, \quad \tau_2 = b/b', \quad \tau_3 = b/a'. \tag{30}$$

For the classical case, i.e., when surface stress is absent, we take $\sigma = \Gamma = \rho_0 = 0$, so that $C = D = 0$, (28) reduces to $\Delta(s) = 0$ which is equivalent to the classical Stoneley wave velocity equation. It is observed that the Stoneley waves under the influence of surface stress are dispersive in contrast with the classical situation for which waves of all frequencies propagate uniformly at the same speed. The equation $\Delta(s) = 0$ was solved numerically by Koppe (1948) who concluded that the velocity of the interface wave falls between the values of the velocity of Rayleigh waves and of the transverse waves in the medium with greater acoustic density. Taking $\mu' \to 0$ in (28) we get the Rayleigh wave velocity equation under the influence of surface stress as

$$\eta^2 C D (1 - ST) + s^2 (CT + DS)\eta - g(s) = 0, \tag{32}$$

where

$$g(s) = (2 - s^2) - 4ST. \tag{33}$$

This result is in agreement with the result obtained in (18) and also that by Murdoch (1976). Further, when there is no surface stress, this equation reduces to the classical Rayleigh wave velocity equation $g(s) = 0$. This equation has one real root $s = s_0$, in the interval $(0, 1)$. If $s = s_1$ satisfies the equation $\Delta(s) = 0$, then $s_0 < s_1 < b\times$ transverse wave velocity in the medium with greater acoustical density. From (28) when $C D \neq 0$ we have

$$\eta = \left\{ -s^2 \left[(CT + DS)(1 - S'T') + \frac{\mu'}{\mu} \tau_2^2 (CT' + DS')(1 - ST) \right] \right.$$

$$\pm \left[s^4 \left[(CT + DS)(1 - S'T') + \frac{\mu'}{\mu} \tau_2^2 (CT' + DS')(1 - ST) \right]^2 \right.$$

$$\left. \left. + 4\Delta(s)CD(1 - ST)(1 - S'T') \right]^{1/2} \right\} \div \{2CD(1 - ST)(1 - S'T')\}.$$

$$\tag{34}$$

If $C = 0$, i.e. in the absence of residual surface tension for a specific value of s we have from (28)

Again when $D = 0$ i.e. in the absence of surface tension at some value of s we have

$$\eta = \Delta(s) \bigg/ \left[s^2 \left\{ CT(1 - S'T') + \frac{\mu'}{\mu} \tau_2^2 CT'(1 - ST) \right\} \right]. \tag{36}$$

If both C and D vanish at some value s', we have $\Delta(s') = 0$ so that Stoneley waves exist. Otherwise there is no propagation of Stoneley waves. The case when σ, $\Gamma > \rho_0 b^2$, C and D will remain positive in $(0, 1)$. Since η is positive, $C < b$, $\tau_1, \tau_2, \tau_3 < 1$ we discard the solution of (34) which is negative for some s in $(0, 1)$. Thus we see that Stoneley wave velocity under the influence of surface stresses may be less than or greater than the classical Stoneley wave velocity but always remains greater than the Rayleigh wave velocity. Moreover, in the case when $0 < \sigma < \Gamma < \rho_0 b^2$, C and D will vanish exactly at one point {each on $(0, 1)$}, say s_2 and s_3 on $(0, 1)$, so that $s_2^2 = \sigma/\rho_0 b^2$, $s_3^2 = \Gamma/\rho_0 b^2$ and hence $0 < s_2 < s_3 < 1$, where C is positive on $(0, s_1)$ and negative on $(s_1, 1)$ and D is positive on (O, s_2) and negative on $(s_2, 1)$ Again when σ, Γ lie in $(0, \rho_0 b^2)$, $CT + DS$ and $CT' + DS'$ vanish at say s_4 and s_5 in (s_2, s_3). We observe that $CT + DS$ is positive on $(0, s_4)$ and negative on $(s_4, 1)$ while $CT' + DS'$ is positive on $(0, s_5)$ and negative on $(s_5, 1)$. Thus the number of positive values of η which satisfy (28) for a particular value of s depends upon the location of s_1 related to s_2, s_3, s_4, and s_5, as well as on the value of $Q(s)$ and zeros of $Q(s)$ where

$$Q(s) = s^4 \left\{ (CT + DS)(1 - S'T') + \frac{\mu'}{\mu} \tau_2^2 (CT' + DS')(1 - ST) \right.$$
$$\left. + 4\, \Delta(s)CD(1 - ST)(1 - S'T') \right\}. \tag{37}$$

Further discussion in the light of Murdoch (1976) is not pursued in this paper due to its complicated nature and cumbersome calculation.

5. Conclusion

Stoneley waves under the influence of surface stresses may be discussed in a manner similar to Rayleigh waves. It is seen that the propagation of Stoneley waves under the influence of surface stresses is not always possible. Under certain restricted conditions such a propagation is possible. Moreover, Stoneley wave velocity under the influence of surface stresses may be less or greater than the classical Stoneley wave velocity, but always remains greater than the Rayleigh wave velocity. We point out here that Stoneley waves are dispersive in character in contrast with the classical situation. Rayleigh waves have been deduced as a particular case of Stoneley waves. In case of Love waves we have shown that the wave velocity equation does not depend on residual surface tension. Some numerical calculations highlight the effect of surface stresses on Love wave as described in the paper.

References

Bullen K E, Bolt 1985 *An introduction to the theory of seismology*, 4th edn (Cambridge: University Press)

Chandrasekharaiah D S 1987 Effect of surface stresses and voids on Rayleigh waves in an elastic solid. *Int. J. Eng. Sci.* 25: 205–211

Chandrasekharaiah D S 1986 Surface waves in an elastic half-space with voids. *Acta Mech.* 62: 77–85

Das S C, Acharya D P, Sengupta P R 1992 Surface waves in an inhomogeneous elastic medium under the influence of gravity. *Mech. Appl.* 37: 541–551

Dey S K, Sengupta P R 1978 Effects of anisotropy on surface waves under the influence of gravity *Acta Geophys. Polon.* 26: 291–298

Gurtin M E, Murdoch A I 1974–75 A continuum theory of elastic material surfaces. *Arch. Rat. Mech. Anal.* 57: 291–323

Gurtin M E, Murdoch A I 1976 Effect of surface stresses on wave propagation in solids. *J. Appl. Phys.* 47: 4414–4421

Koppe H 1948 Über Rayleigh–Wellen ander oberflächezweier Medien. *Z. Angew. Math. Mech.* 28: 355–360

Murdoch A I 1976 The propagation of surface waves in bodies with material boundaries. *J. Mech. Phys. Solids* 24: 137–146

Pal P K, Acharya D P, Sengupta P R 1996 Effect of voids on the propagation of waves in an elastic layer. *Sadhana* 21: 477–485

Plaster H J 1972 *Blast cleaning and allied processes* (London: Industrial Newspapers)

Recent Advances in Power Electronics and Drives

Foreword

Due to availability of self-commutated devices like power MOSFET, IGBT, and GTO, power electronics has undergone substantial change in recent years. It is used now in a very large number of applications. Some of them are: VAR compensators, active power filters, switch mode rectifiers, space and uninterruptible power supplies, industrial fans, blowers, compressors, steel rolling mills, paper and textile machines, mine winders, cement mills, electric vehicles, electric and diesel electric traction, high voltage *dc* transmission, induction heating, electrolysis, excavators, ship propulsion, missiles, satellite tracking stations, machine tools, robots, lasers, domestic appliances, defence electronics, flexible *ac* transmission etc.

While the applications of power electronics are growing, the main emphasis of R & D has shifted to improvement of efficiency and performance, reliability, and power quality, reduction in volume and weight, and in maintenance and down-time, longer life, fast/optimal response and mitigation of problems associated with harmonics and poor power factor.

This issue of *Sādhanā* aims at making the state-of-the-art reviews of the emerging topics available to its readers. Papers in the issue have been contributed by eminent researchers who are internationally known for their contributions in the respective areas.

Performance and efficiency of converters, and associated drives and power systems, depend to a great extent on the modulation technique. Because of simple digital realisation, flexibility in adaptation to various applications and good performance, space vector pulse width modulation has become quite popular. V T Ranganathan provides a status review of space vector pulse-width modulation in the first article.

Simulation permits detailed analysis leading to deeper insight into the behaviour of power electronic systems. Consequently one can select (or design) optimal parameters and control techniques, and then proceed to build the systems. Therefore, the accepted practice today is to first carry out detailed simulation study using the simulation packages developed particularly for power converters. The paper by M Dawande, Victor Donescu, Z Yao and V Rajagopalan reports the recent advances in simulation of power electronic converter systems.

Static VAR compensators have been for many years an essential component in the operation of power transmission and distribution systems. While in transmission systems they support the line voltage and stabilise the system, in the case of distribution systems, they allow improvement of power factor and provide support to line voltage. Availability of self-commutated devices has led to a new generation of VAR compensators known as

Modern VAR compensators or Synchronous Link Converter Based VAR Compensators or simply Static Compensators (STATCOM). They have the advantages of compact size, high efficiency, low harmonic content and stepless control of VARs from full lagging to full leading. The paper by Geza Joos on recent advances in VAR compensators covers compensators based on current and voltage source converters and on *ac* controllers, both in shunt and series configuration.

The combined capacity of loads which pollute the power system with harmonics has reached intolerable levels, making it necessary to take measures to reduce the pollution to an acceptable level so that the quality of power supply can be preserved. Several of these loads consist of systems with thyristor-based converters as front-end converters, which generate harmonics and operate at low power factor. This problem of power system pollution is being tackled by two measures. First, a new generation of converters is being developed which will operate at unity power factor and draw sinusoidal current from the mains. This will substantially eliminate the pollution caused by converters. The second measure consists of evolution of active power filters to filter out harmonics. Unlike passive power filters, active power filters have compact size, higher efficiency, lower cost and better ability to filter harmonics. These issues of great concern to the present day power system have been tackled in three papers. The paper by Ned Mohan and G R Kamath on the recent advances in active power filters describes some unity power factor and sinusoidal current converters and the principle of active power filters. The advantages of resonant converters are well established. The paper by A K S Bhat on high power factor operation of resonant converters presents the operation and characteristics of series-parallel (LCC-type) resonant converters operating in discontinuous and continuous current modes with high power factor and low harmonic current distortion for wide variation of load. The review of single phase power factor correction by Ramesh Oruganti and Ramesh Srinivasan covers 1-phase *ac–dc* converters capable of operating at a high power factor and with harmonic distortion in source current within acceptable limits.

The desire to transmit more power through a given transmission and distribution network and at the same time maintain the quality of supply has led to the development of a new area of research known as flexible *ac* transmission system. It employs power electronic converters. It is projected that by 2010 around 80% of electric power will flow through power electronic converters. K R Padiyar and A M Kulkarni present a status review of flexible *ac* transmission systems in their paper and provide a generalised description of FACT (flexible *ac* transmission) controllers and a review of progress in FACT systems.

Another major application of power electronics is in the control of electric drives. In several applications, electric drives must have a fast torque response, four quadrant operation capability and controllability of torque and speed over a wide range of operating conditions. Traditionally *dc* motor drives have been used in these applications; *ac* motor drives have several advantages over *dc* drives. With vector control they are able to fulfill these requirements very well, hence vector controlled *ac* drives find wide application. Though vector control by now has gained maturity it continues to be a topic of great interest to researchers. The review paper by A K Chattopadhyay discusses advances in vector control of *ac* motor drives and describes various aspects of vector control of induction motor and synchronous motor drives.

Switched reluctance motors have several advantages over induction and synchronous motors. Prominent among these is simple and robust construction. Switched reluctance motor drives are therefore projected to be cheaper and more sturdy. Several problems are still to be solved before a commercially acceptable general purpose drive can be developed. The review of recent advances in switched reluctance motor drives by M Ehsani discusses various aspects of switched reluctance motor drives.

Due to the advance in magnetic material technology and control algorithms, the applications of permanent magnet brushless dc motors are growing at a fast rate, particularly in low power applications. The paper by Bhim Singh discusses recent advances in permanent magnet brushless dc motors both with rectangular and sinusoidal fields.

The Indian Railways are in the process of modernising locomotives and EMU drives. Their ultimate aim is to replace dc drives by ac motor drives. Being a specialised topic, literature covering all the aspects is not easily available. In their status review on ac motor traction drives L Frederick and G K Dubey cover ac motor traction drives.

It is hoped that this issue of *Sādhanā* will provide its readers with a good exposure to recent advances in several topics of current research interest.

On behalf of the Indian Academy of Sciences and on my own behalf, I wish to express sincere thanks to all contributors to this issue.

I also wish to record my sincere thanks to Prof. N Viswanadhan, editor of *Sādhanā*, for giving me an opportunity to edit this special issue. I also thank Prof. M A Pai of the University of Illinois for encouragement and help.

Ms. Shashikala and her team at the Academy also deserve appreciation for working with interest and enthusiasm to bring out the issue in good shape.

December 1997

GOPAL K DUBEY
Guest Editor

Space vector pulsewidth modulation – A status review

V T RANGANATHAN

Department of Electrical Engineering, Indian Institute of Science,
Bangalore 560 012, India
e-mail: vtran@ee.iisc.ernet.in

Abstract. The technique of space vector pulsewidth modulation (SVM) is reviewed. The basic principle of SVM is derived and is compared with sine-triangle PWM. Operation in the overmodulation range is explained. Extension of SVM to other inverter-motor combinations such as three level inverters and split phase motors are discussed.

Keywords. Pulsewidth modulation (PWM); space vectors; space vector modulation (SVM).

1. Introduction

A large number of variable speed induction motor drives are in use today in applications as varied as pumps and fans, traction, machine tool spindles, printing presses etc. The vast majority of them are based on voltage source inverters, although other circuit configurations are possible. This is because the developments in power device technology have resulted in the availability of a number of power switches, such as MOSFET, IGBT and GTO, which meet every kind of power rating, from a few hundred watts to megawatts. Voltage source inverters, as is well known, produce pulsed output waveforms, consisting of the fundamental component which is required to drive the motor, as well as the harmonics which contribute only to the losses and torque pulsations. The method of determining the widths and the sequence of the voltage pulses produced by the inverter is known as Pulsewidth Modulation (PWM).

The basic attempt in all Pulsewidth Modulation techniques is to produce the required amplitude and frequency of the fundamental voltage necessary for driving the motor, while moving the energy in the harmonics to a higher range in the frequency spectrum. The impedance presented by the motor to these harmonic voltages consists mainly of the leakage reactance. The expectation is that at such higher frequencies, even the reactance of the machine leakage inductance will be appreciable, thereby limiting the harmonic currents drawn from the inverter. Also, since the torque pulsations created by the harmonic currents will also be at a higher frequency, the motor should be able to run smoothly.

The constraint on the PWM process is the fact that additional switchings per cycle are required in order to accomplish the modulation. This increases the losses in the inverter. Moreover, as switching patterns become more complicated, an individual inverter phase may be required to produce pulses or notches of very small width. The swiching times of the power devices used in the inverter impose a limitation here.

Several popular PWM techniques have emerged over the years. The earliest one was the sine triangle method (Schonung & Stemmler 1964), in which the switching instants of the inverter legs are decided by the points of intersection of a high frequency (anywhere from 450 hertz to a few kilohertz) triangular wave and a sinewave at the frequency of the required fundamental motor voltage. The triangle is referred to as the carrier and the sine as the reference. The output voltage produced by the inverter contains the fundamental component, of which the sine wave is a replica. In addition, it contains harmonic components which are located in bands around multiples of the triangle frequency. It is, in general, necessary to synchronise the carrier to the sine wave, in order to avoid frequency components in the output below the sine frequency – so called subharmonic components.

The sine triangle method of PWM, though used widely, has its limitations. The amplitude of the sine wave has to be increased in proportion to its frequency, in order to keep the motor flux constant. Therefore, at higher motor speeds, a situation arises where the amplitude of the sine wave becomes equal to that of the triangle. If the sine wave amplitude were to increase any further, the linearity between the sine and the fundamental component of the inverter output voltage is lost. This situation is referred to as overmodulation. A further consequence of overmodulation is that frequency components with low harmonic numbers (5th, 7th etc.) begin to appear in the output voltage. From the implementation point of view also, sine triangle PWM requires some care. Where the triangle frequency is limited by the switching speed of the power devices, analog realisations give the best results in terms of output waveform quality. However, analog realisations require considerable ingenuity, especially to achieve synchronisation (Kliman & Plunkett 1979; Green & Boys 1982).

Subsequently, other techniques such as Selected Harmonic Elimination (Pollmann 1983), which were more amenable to digital implementation, have also been used widely. Today, the trend being towards standard hardware based on digital processors, it is preferred to implement all PWM techniques through real time software execution of algorithms. A number of software realisations of the sine triangle technique have been reported in the literature. They are good approximations of the analog realisation if the frequency of the triangle is high (at least 1 kilohertz or above). The limitation of the output voltage due to overmodulation effects is, however, still a factor to be considered.

The above techniques have one thing in common in that the PWM patterns are calculated for each phase, independent of switchings in the other phases. It has subsequently been realised that good results can be obtained if the combined effect of all the three output voltages of the inverter on the motor are taken into account in generating the switching patterns. The conceptual framework for doing so is the space vector modelling of *ac* motors. In the following sections, the notion of space vectors is introduced and the technique of generating PWM patterns based on these viz. space vector PWM is reviewed.
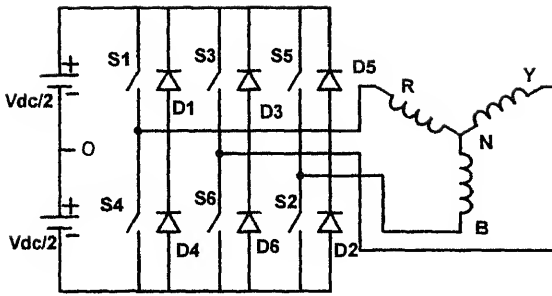
Figure 1. Three-phase half-bridge voltage source inverter.

2. Basic space vector PWM (Holtz *et al* 1987; Van der Broek *et al* 1988)

2.1 *Space vectors*

Consider a three-phase half-bridge voltage source inverter, shown in figure 1. Each phase to centre-tap voltage , viz. V_{RO}, V_{YO}, V_{BO}, can have only two possible values, namely $+V_{dc}/2$ or $-V_{dc}/2$ respectively. As there are three switches, corresponding to the three phases, there are eight possible states for the inverter at any instant of time. Corresponding to each of the switching states, the motor line to neutral voltages can be determined using the following equations,

$$
\begin{aligned}
V_{s1} &= V_{RN} = (1/3)(V_{RY} - V_{BR}),\\
V_{s1} &= (1/3)(2V_{RO} - V_{YO} - V_{BO}),\\
V_{s2} &= (1/3)(2V_{YO} - V_{BO} - V_{RO}),\\
V_{s3} &= (1/3)(2V_{BO} - V_{RO} - V_{YO}).
\end{aligned}
\tag{1}
$$

From these three-phase voltages, equivalent two-phase voltages are defined as follows,

$$
\begin{aligned}
V_{sa} &= (3/2)V_{s1},\\
V_{sb} &= (\sqrt{3}/2)(V_{s2} - V_{s3}).
\end{aligned}
\tag{2}
$$

(1) + - -   (2) + + -   (3) - + -   (4) - + +

(5) - - +   (6) + - +   (7) + + +   (8) - - -
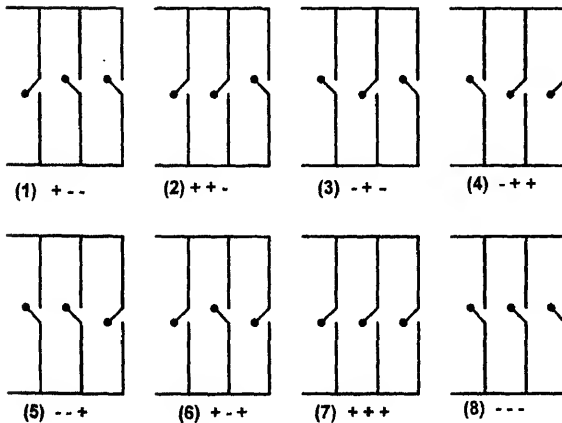
Figure 2. Eight switching states of the inverter.
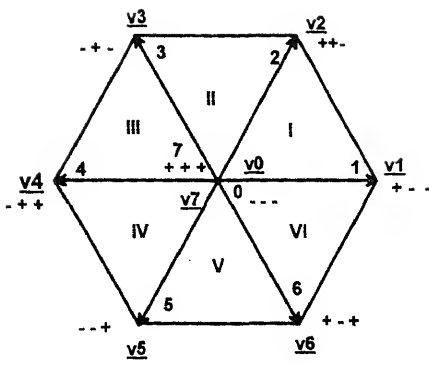
Figure 3. Space vector corresponding to each switching state.

The space vector of the machine stator voltage is then defined as

$$V_s = V_{sa} + jV_{sb}. \tag{3}$$

The eight switching states of the inverter and the corresponding space vectors of the machine stator voltage are shown in figures 2 and 3 respectively.

It can be seen that at any instant of time, there are only eight possible positions for the voltage space vector. When the inverter is operated in the six-step mode, the switching states go through the sequence 1-2-3-4-5-6-1-2-3. The states 0 and 8 are not used at all. The space vector of stator voltage stays in each of the positions 1 to 6 for a time interval corresponding to 60° of the fundamental period and jumps to the next position at the end of every sixty degrees (one-sixth of the fundamental period). With above switching sequence, at every jump in the position of the voltage space vector, only one of the three phases of the inverter switches from top to bottom or vice versa.

2.2 *Space vector PWM*

The ideal trajectory for the voltage space vector is of course a circle described with uniform angular velocity, which results only when the motor is fed from a three-phase sinusoidal voltage source. The objective of any PWM process is therefore to approximate this ideal trajectory of the voltage space vector by switching amongst the eight standard positions. Towards this end, the continuously moving reference vector $\mathbf{V_s^*}$ is sampled at a sampling frequency f_s. During the interval $T_s = 1/f_s$ between samples, the reference vector is assumed to remain constant. It is clear that for this assumption to be valid, the sampling frequncy should be fairly high compared to the fundamental output frequency desired from the inverter.

Now consider that at a particular sampling instant, the reference $\mathbf{V_s^*}$ is situated in sector I as shown in figure 4. The angle α represents the position of the reference vector with respect to the beginning of the sector. It is intuitively clear that the reference vector can be reproduced best during the period till the next sample by switching the inverter to create the vectors $\mathbf{V_1}$, $\mathbf{V_2}$, $\mathbf{V_0}$ and $\mathbf{V_7}$ in some sequence. Selecting any of the other vectors would result in a greater deviation of the actual vector from the desired reference and would thus contribute to harmonics.
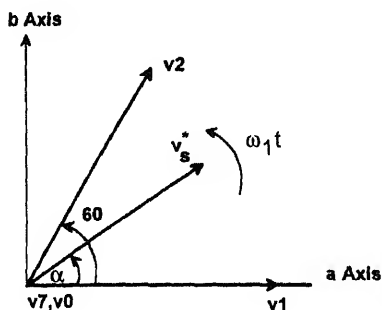
Figure 4. Sampled reference vector located in sector I.

The switching pattern can be calculated as follows. Assume that the sampling period T_s is divided into three subintervals T_1, T_2 and T_0. The inverter is switched so as to produce the vector $\mathbf{V}_1$ for T_1 seconds, $\mathbf{V}_2$ for T_2 seconds and zero (either $\mathbf{V}_0$ or $\mathbf{V}_7$) for T_0 seconds. The subintervals have to be calculated so that the volt-seconds produced by these vectors along the a and b axes are the same as those produced by the desired reference vector $\mathbf{V}_s^*$ i.e.

$$V_{dc} \cdot T_1 + V_{dc} \cdot \cos 60 \cdot T_2 = |\mathbf{V}_s^*| \cdot \cos \alpha \cdot T_s \cdots a \text{ axis}, \tag{4}$$

$$V_{dc} \cdot \sin 60 \cdot T_2 = |\mathbf{V}_s^*| \cdot \sin \alpha \cdot T_s \cdots b \text{ axis}, \tag{5}$$

where $|\mathbf{V}_s^*|$ is the amplitude or the length of the reference vector. Define the amplitude ratio

$$a = |\mathbf{V}_s^*| / V_{dc}. \tag{6}$$

Then (4) and (5) can be rewritten as,

$$T_1 + T_2 \cdot \cos 60 = a \cdot T_s \cdot \cos \alpha, \tag{7}$$

$$T_2 \cdot \sin 60 = a \cdot T_s \cdot \sin \alpha. \tag{8}$$

Solving for T_1 and T_2,

$$T_1 = T_s \cdot a \cdot \sin(60 - \alpha) / \sin 60, \tag{9}$$

$$T_2 = T_s \cdot a \cdot \sin \alpha / \sin 60, \tag{10}$$

$$T_0 = T_s - T_1 - T_2. \tag{11}$$

As the reference vector moves to other sectors, the corresponding boundary vectors of the sector should be created during the intervals T_1 and T_2.

In order to minimise the number of switchings in the inverter, it is desirable that switching should take place in one phase of the inverter only for transition from one state to another. This can be achieved if the following switching sequence is used:

$$0\text{-}1\text{-}2\text{-}7|\text{-}7\text{-}2\text{-}1\text{-}0|\text{-}0\text{-}1\text{-}2 \dots .$$

Therefore the zero interval is divided into two equal halves of length $T_0/2$. These half-
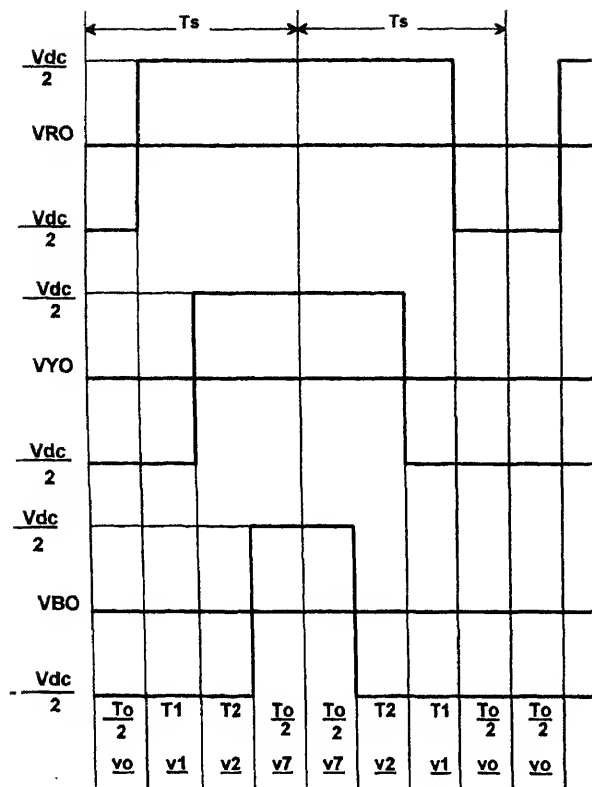
Figure 5.   Inverter phase to centre tap voltages.

sequence during successive sampling intervals will then be as shown in figure 5. It can be seen that the chopping frequency f_c of each phase of the inverter is given by:

$$f_c = f_s/2. \tag{12}$$

It should be pointed out that during the sampling interval, the desired reference vector is approximated in the average sense, since the volt-seconds are equated. However, instantaneously, the actual vectors produced by the inverter are different from the reference vector and therefore instantaneous voltage deviations or voltage 'ripple' exists. As a result harmonic currents will also flow in the machine. By following the above mentioned switching sequence, harmonics are to some extent reduced.

2.3   Interpretation in terms of sine-triangle PWM

In order to compare the space vector PWM with sine-triangle PWM, the mean value of the inverter phase to the *dc* centre tap voltages are calculated first, over one full cycle of the fundamental i.e. for one full rotation of the reference vector $\mathbf{V}_s^*$. Let $t = 0$ correspond to the instant where the average value V_{RO} of the phase to centre tap voltage goes through its positive zero crossing. At this instant, therefore, the reference vector will point vertically downwards as shown in figure 6. From this initial position, the reference vector rotates in the anticlockwise direction with angular velocity ω_1. Its position at any instant t with
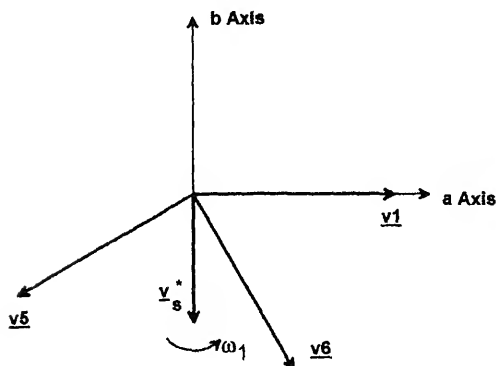
Figure 6. Reference time instant for calculating V_{RO}.

respect to the starting position is given by the angle $\theta = \omega_1 t$. For $0 \leq \theta \leq 30°$, the reference vector is in sector V; for $30 \leq \theta \leq 90°$, the reference vector lies in sector VI. If the variation of V_{RO} can be determined over these 90°, then its waveform over the rest of the cycle can be drawn using symmetry.

$$\text{For } 0 \leq \theta \leq 30°, \quad V_{RO} = (V_{dc}/2)(1/T_s)[-T_1 + T_2]. \tag{13}$$

Further $\alpha = (30 + \theta)°$. Using the expressions (9) and (10) for T_1 and T_2, it can be shown that

$$V_{RO} = V_{dc} \cdot a \cdot \sin \theta = V_{dc} \cdot a \cdot \sin \omega_1 t. \tag{14}$$

$$\text{For } 30 \leq \theta \leq 90°, \quad V_{RO} = (V_{dc}/2)(1/T_s)[-T_1 + T_2]. \tag{15}$$

However, $\alpha = \theta - 30°$. Again, using (9) and (10) in (15), it can be shown that

$$V_{RO} = (V_{dc} \cdot (1/\sqrt{3}) \cdot a \cdot \sin(30 + \omega_1 t). \tag{16}$$

Using the expressions in (14) and (16) and taking advantage of symmetry, the variation of V_{RO} over one full cycle of the fundamental can be drawn. The waveforms for the other two phases will be identical with phase shifts of 120°. The mean value of the line voltages can also then be drawn by taking the differences. Figure 7 below shows the variations over the positive half cycle of phase R.
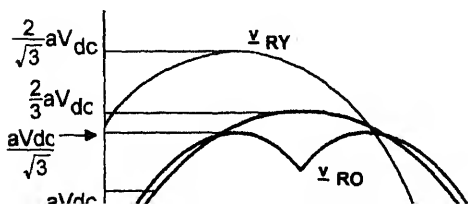
From the figure, it can be seen that the waveform of the mean value of the phase to centre tap voltage is not sinusoidal but contains triplen harmonics. These harmonics get cancelled in the line to line voltages, which are therefore sinusoidal. The machine line to neutral voltages will also be sinusoidal. Considering the waveform of the line to line voltage in figure 7,

$$\text{Peak value of mean line to line voltage} = (2/\sqrt{3}) \cdot a \cdot V_{dc}. \tag{17}$$

Recalling the definition of the voltage control ratio given by (6), the maximum value of a is given by

$$a_{max} = \text{radius of the largest circle inscribed in the hexagon} \div V_{dc}. \tag{18}$$

Therefore,

$$a_{max} = (\sqrt{3}/2). \tag{19}$$

Substituting in (17), the maximum peak line to line voltage using SVM

$$= (2/\sqrt{3}) \cdot (\sqrt{3}/2) \cdot V_{dc} = V_{dc}. \tag{20}$$

The corresponding value for sine-triangle modulation is given by $(V_{dc}/2) \cdot \sqrt{3}$. It can therefore be concluded that SVM can produce 15% more fundamental compared to the sine-triangle method without going into over-modulation. SVM can be regarded as a carrier-based technique, with the modification that the reference waveforms have triple harmonics added to them. In fact, modulators have been designed with a triangular waveform at three times the sine frequency being added to the sine references.

2.4 Comparison with sine-triangle PWM in terms of waveform quality

In general, comparison of the waveforms produced by different PWM techniques is carried out in terms of the 'loss factor' (Handley & Boys 1992; Holtz 1994). These are measures of the total *rms* harmonic currents flowing in the motor, usually normalised with respect to the harmonics produced by a six-step inverter (Holtz 1994). A comparison of SVM with sine-triagle in this respect shows that SVM produces less current harmonics than sine-triangle at higher modulation indices. At low modulation indices (< 0.4), there is no marked superiority of one technique over the other.

3. Overmodulation

As the fundamental output frequency of the inverter increases, and with it the amplitude of the reference space vector $\mathbf{V_s^*}$, the duration of time t_0 to be spent at the zero vector becomes smaller and smaller. When the locus of the reference vector becomes the circle inscribed within the hexagon, t_0 becomes zero. This is spoken of as the limit of modulation. For further increase in the length of $\mathbf{V_s^*}$, the locus lies partly outside the hexagon, implying that for some of the samples, the reference vector cannot be approximated in the volt-second sense by the SVM technique, even if the duration of t_0 is made zero. Different policies can be adopted to generate the pulse patterns for such loci. One technique is to apply equations
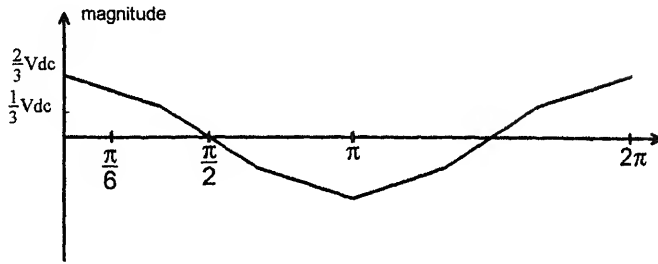
Figure 8. Average phase to centre tap voltage for overmodulation.

(9) to (11) for those α values for which the locus lies inside the hexagon. For other values of α, the circular locus is forsaken in favor of the hexagon. This can be achieved by setting:

$$t_0 = 0,$$

$$t_1' = t_1/(t_1 + t_2); \ \text{ i.e. } \ t_1' = (T_s) \frac{\sqrt{3}\cos\alpha - \sin\alpha}{\sqrt{3}\cos\alpha + \sin\alpha}, \tag{21}$$

$$t_2' = T_s - t_1'.$$

This procedure scales down the subintervals of the non-zero vectors so that they add up to the sampling interval T_s, while setting the subinterval for the zero vector to zero. Therefore, there is an error in the amplitude of the motor voltage for such samples. The limit of this process is reached when the circular locus of the reference circumscribes the hexagon. The average phase voltage waveform of the motor will then have the waveform shown in figure 8. The amplitude of the fundamental phase voltage is

$$\frac{9}{\pi^2} \cdot \frac{2}{3} \cdot V_{dc} = \frac{6}{\pi^2} \cdot V_{dc}. \tag{22}$$

The voltage of figure 8 contains low order harmonics such as the 5th and the 7th.

4. Modified SVM or bus clamped SVM

In conventional SVM, there are three switchings in the inverter in each sampling interval e.g. V_0, V_1, V_2 and V_7. If it is desired to reduce the number of switchings, an alternative sequence such as V_0-V_1-$V_2|V_2$-V_1-V_0 can be thought of. Here the entire time for the zero vector is spent only at V_0. Compared to the conventional SVM, the number of switchings in the inverter becoms $2/3$ in this case. The effect of such a switching strategy can be analysed along lines similar to that in § 2.1. The average pole voltage waveform can be shown to have a time variation of the form indicated in figure 9. This process can therefore
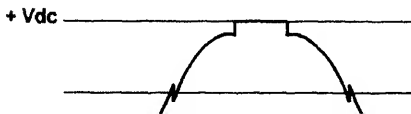
be considered conceptually equal to a sine-triangle PWM , where a rectangular waveform at three times the sine frequency is added to each of the sine waves. Since one of the inverter phases is tied to a bus over 60° intervals, this technique has also been referred to as Bus Clamped SVM. This technique has the advantage that the lowest order harmonic generated is at a frequency 50% higher than that of normal SVM with the same switching rate. In addition, since there are only four swichings over a PWM period, there is less overhead in terms of loading timers, when it comes to a question of practical realisation. However, the performance in terms of harmonic content or loss factor is inferior to conventional SVM.

5. Extensions of space vector modulation

The majority of drives use a single inverter with a normal three phase *ac* motor. In high power drives, however, in order to limit the voltage and current ratings of the power devices, other configurations are also in use. Notable amongst these are: (i) the three-level inverter (Nabae *et al* 1981; Stemmler 1994) and (ii) the split phase induction motor drive (Andresen 1981; Gopakumar *et al* 1993). These configurations extend the power rating of the system by using double the number of switches i.e.twelve. Space vector concepts can be extended to such configurations also in order to generate the pulse patterns, as outlined below.

5.1 *SVM of three level inverter*

The configuration of a three level inverter is schematically indicated in figure 10. Each phase to centre tap voltage can now have three values viz.$+V_{dc}/2, 0, -V_{dc}/2$. By considering all the possible switching states and applying the definition of voltage space phasor in (3), it can be shown that the possible locations for the voltage space vector are as shown in figure 11.

Although theoretically there are $3^3 = 27$ possible space vector locations, there are only 18 distinct vectors in addition to the zero vector. Vectors 13 to 18 can each be realised by more than one switching state of the inverter. The zero vector can also be realised by three different switching states, namely $(+ + +)$, (000), and $(- - -)$.
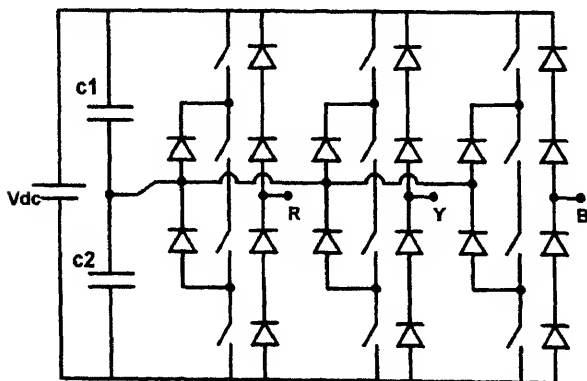
Figure 10. Circuit diagram of three-level inverter.
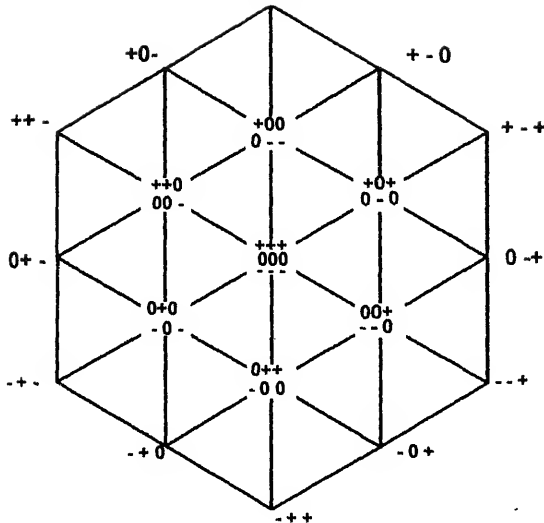
Figure 11. Voltage space vector for three-level inverter.

A strategy for generating the switching sequence can again be arrived at by looking at a 60° segment, as shown in figure 12 below (Kolenchery *et al* 1996). The sector can be divided into 4 triangles. Depending on the triangle in which the tip of the reference vector lies, the vectors and the associated switching states can be judiciously selected so as to approximate the reference in the volt-seconds sense as well as to keep the number of switchings in the inverter low.

5.2 SVM for split phase motors

An alternative to the three-level inverter for high power drives is the split-phase motor (Andresen 1981; Gopakumar *et al* 1993). The split phase motor is obtained by splitting the phase belt of a conventional three-phase motor into two equal halves with a phase separation of 30° between the two halves (figure 13). The split phase groups are controlled by two inverters with a dc link voltage of $V_{dc}/2 \cos 15$ each. Since the split phase configuration
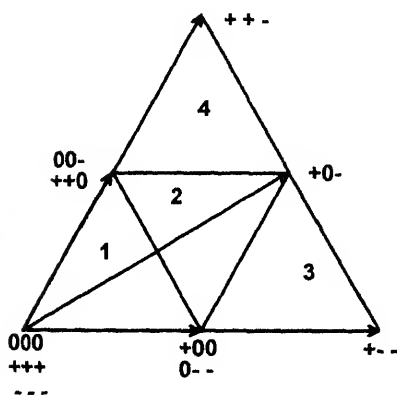
Figure 12. 60° sector of space vector hexagon for three-level inverter.
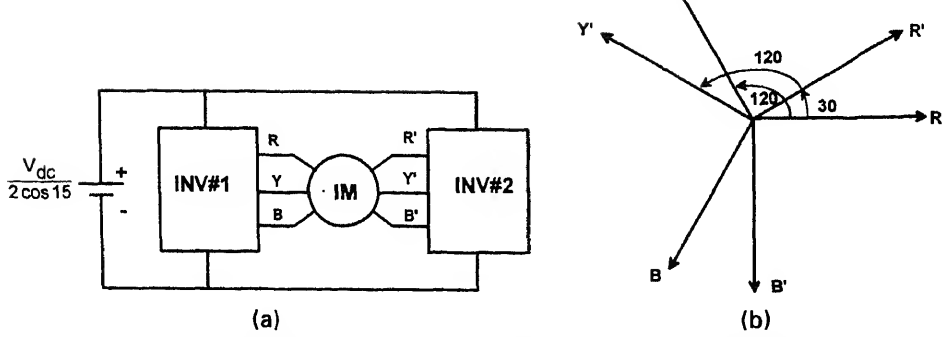
Figure 13. Configuration of split-phase induction motor drive. (a) Block diagram of split-phase induction motor drive. (b) Orientation of MMF vectors produced by the coils of a split-phase machine.

is achieved by splitting the phase belts in two, the equivalent number of turns per phase for the new configuration is $N_s/2 \cos 15$, where N_s is the equivalent number of turns for the three-phase configuration. Correspondingly, a link voltage of $V_{dc}/2 \cos 15$ gives the same magnitude for the airgap flux in the new configuration as compared to the three phase inverter drive with a dc link voltage of V_{dc}.

The voltage space vector locations for the two sets of coils are shown in figure 14. The locations of one group are shifted by 30° from those of the other. There are 48 possible distinct space vector locations by combining the individual locations of the two groups. The boundary of all the locations is a 12-sided polygon, as compared to the hexagon for a single inverter. The reference vector can therefore be located in any one of the 12 sectors. One possible method of generating the pulse patterns for the switches is to approximate the reference vector by the 12 outermost vectors (Gopakumar *et al* 1993). The subintervals T_1, T_2 and T_0 are then obtained as

$$T_1 = 2 \cdot a \cdot T_s \cdot \sin(30 - \alpha),$$
$$T_2 = 2 \cdot a \cdot T_s \cdot \sin \alpha, \tag{23}$$
$$T_0 = T_s - T_1 - T_2,$$
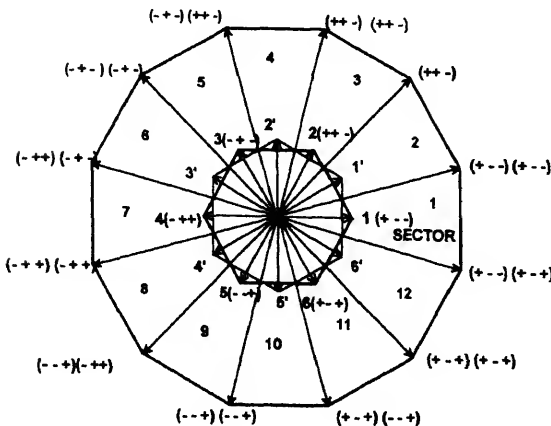
Figure 14. Voltage space vector location for split-phase drive.
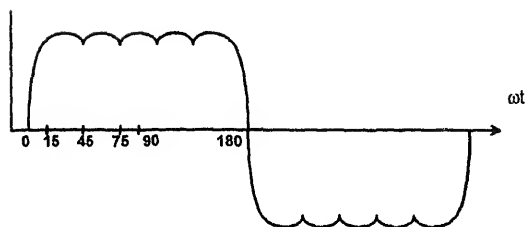
Figure 15. Average phase centre tap voltage for split-phase drive.

where now α is the angle of the reference from the beginning of each 30° sector. By an anlysis similar to that in § 2.3, it can be shown that the average pole voltage has the time variation of figure 15. It can be seen that these contain harmonics such as the 5th and 7th. These harmonics, though they contribute to stator current, do not produce torque harmonics as their fields get cancelled out in the airgap because of the winding arrangement. With such a technique, it can be shown that an equivalent three-phase voltage of up to $0.643\,V_{dc}$ can be generated before entering into the overmodulation region. This can be compared to $0.577\,V_{dc}$ for the normal three-phase system.

6. Conclusion

The concept of space phasors thus provides a unified framework for the generation of pulse patterns for inverters. It can be extended to configurations other than the simple three phase half-bridge inverter. Space phasor-based methods of PW M are also easily amenable to digital realisation. They can also be easily added on to advanced motor control algorithms such as vector control. Due to these reasons, such methods have become popular in industrial drives. Current control techniques such as hysteresis control can also be adapted for execution in the space phasor domain, although it has not been possible to present such methods here.

References

Andresen E 1981 Six-phase induction motors for current source inverter drives. *Conference Record, IEEE-Ind. Appl. Soc. Annual Meeting*, pp 607–618

Gopakumar K, Ranganathan V T, Bhat S R 1993 Split phase induction motor operation from PWM voltage source inverter. *IEEE Trans. Ind. Appl.* IA-29: 927–932

Green R M, Boys J T 1982 PWM sequence selection and optimisation: A novel approach. *IEEE Trans. Ind. Appl.* IA-18: 146–151

Handley P G, Boys J T 1992 Practical real-time PWM modulators: An assessment. *Inst. Elec. Eng. Proc.* B139: 96–102

Holtz J, Lammert P, Lotzkat W 1987 High speed drive system with ultrasonic MOSFET PWM inverter and single-chip microprocessor control. *IEEE Trans. Ind. Appl.* IA-23: 1010–1015

Holtz J 1994 Pulsewidth modulation for electronic power conversion. *Proc. IEEE* 82: 1194-1214

Kliman G B, Plunkett A B 1979 Development of a modulation strategy for a PWM inverter drive. *IEEE Trans. Ind. Appl.* IA-15: 72–79

Kolenchery S S, Vaidya V C, Madhu Mangal 1996 SVM-PWM strategy for high power 3-level inverters in variable frequency applications. *Proc. IEEE Int. Conf. Power Electronic Drive and Energy Systems for Industrial Growth (PEDES'96)* 1: 197–200

Nabae A , Takahashi I, Akagi H 1981 A new neutral-point-clamped PWM inverter. *IEEE Trans. Ind. Appl.* IA-17: 518–523

Pollmann A 1983 A digital puslewidth modulator employing advanced modulation techniques. *IEEE Trans. Ind. Appl.* IA-19: 409–413

Schonung A, Stemmler H 1964 Static frequency changer with subharmonic control in conjunction with reversible variable speed AC drives. *Brown Boveri Rev.* 51: 555–577

Stemmler H 1994 High power industrial drives. *Proc. IEEE* 82: 1266–1286

Van der Broek H W, Skudelni H, Stanke G V 1988 Analysis and realisation of a pulsewidth modulator based on voltage space vectors. *IEEE Trans. Ind. Appl.* IA-24: 142–150

Recent advances in simulation of power electronics converter systems

MANJUSHA DAWANDE*, VICTOR DONESCU, ZIWEN YAO and
V RAJAGOPALAN

CPEE-Hydro-Québec/NSERC Industrial Research Chair, Département de génie
électrique, Université du Québec à Trois-Rivières, C.P.500, Trois-Rivières,
Québec, G9A 5H7, Canada
*Present address: Department of Electrical Engineering, University of Roor-
kee, Roorkee 247 667, India
e-mail: kmsee@rurkiu.ernet.in; rajagopalan@uqtr.uquebec.ca

Abstract. This paper presents a comprehensive state-of-the-art of the recent
advances in simulation of power electronic converter systems. Knowing the
importance of simulation, this paper reviews the various methods of modelling,
circuit analysis approaches, numerical techniques etc. Several general purpose
simulators and dedicated power electronic simulators have been discussed.
A few demonstrative examples of simulation of power electronic converters
by using different simulators are provided. Practical difficulties in simulation,
challenges, new developments and scope for future work are also discussed.

Keywords. Power electronic converter; power electronic simulator.

1. Introduction

Power electronics has gone through intense technological evolution during the last three
decades. Power electronic converter systems have captured a major place in it, as they find
wide applications in industry. It has been reported in the recent IEEE PELS (Newsletter
1997) that 'Analysis and control of power converters' is the area of highest interest (65%)
in power electronics. At present, in research, design and development power electronic
converter systems, modelling and simulation are widely used and became indispensable
before practical hardware implementation.

Design of power electronic converters requires fairly sophisticated analysis, simulation
tools and detailed design procedures involving tolerance analysis and worst-case design
before building the actual system. The computation of steady state solution is an essential
first step in most of the design studies (Rajagopalan 1987). Computer-based design au-
tomation tools are expected to have an increasing influence on improving the reliability of
power electronic systems (Kang & Lavers 1994). Accurate, fast and cheaper simulation

689
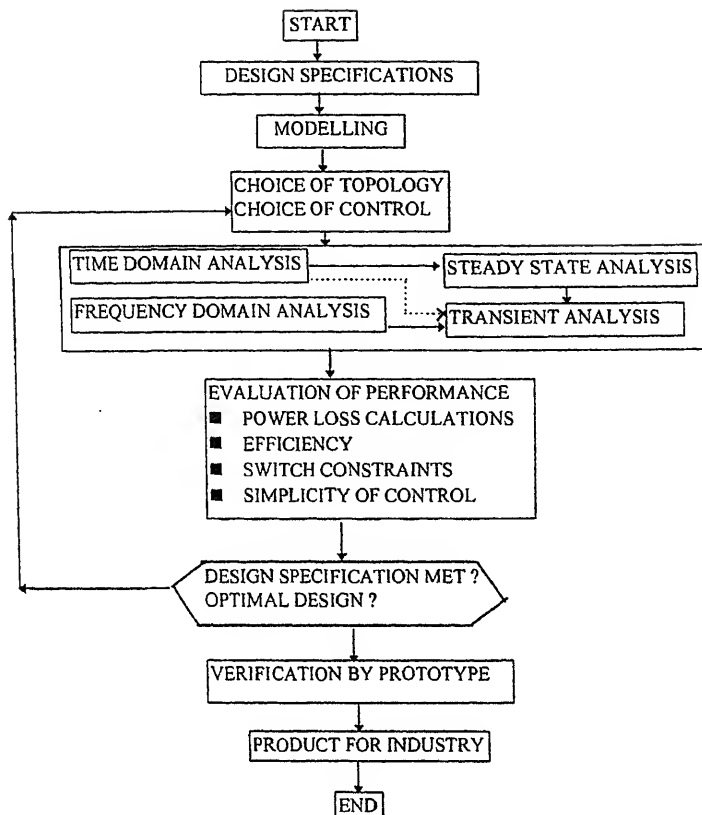
Figure 1. Flow chart for design process.

methods are required for (1) comprehensive system performance, (2) efficient designs, (3) discovering possible problems due to abnormal operating modes, (4) optimal design, (5) sensitive study based on the various parameters, (6) knowing the behavior of circuit at various levels, (7) predicting the interaction with other networks, and (8) avoiding destructive laboratory tests etc. Therefore, the normal design process should follow the algorithm given in figure 1. It is always advisable to study the system thoroughly, to determine the influence of all the parameters, to raise the maximum doubts for its worst-case operation and get the solutions through simulation, before finalizing the optimal design of power electronic converters. Micro-modelling and macro-modelling of power semiconductor devices, new simulation tools and user-friendly environments are currently available for the simulation of power electronic converters (Rajagopalan *et al* 1990, 1992, 1995; Rajagopalan 1995).

This paper gives a comprehensive state-of-art of the several simulation techniques employed to the power converters. It describes the micro- and macro-modelling of power components, time and frequency domain analysis, different circuit analysis approaches, various numerical techniques used in simulation etc. Several general purpose simulators such as MATLAB-SIMULINK, SIMNON, EMTP, EMTDC, PSPICE, SABER etc. are reviewed. Several dedicated power electronic circuit simulators such as ATOSEC5, PSIM, SIMUPELS are briefly discussed. Computer-aided learning tools like CATPELS are also

mentioned. Some examples are demonstrated in different simulators. The paper discusses some of the practical difficulties in the simulation process and explores the challenges that are to be overcome to make simulation more efficient and attractive to design engineers. Finally, the future trend of systematic controller design, i.e. hardware implementation directly from digital simulation, is introduced.

2. Simulation techniques

The importance of simulation is wisely referred to as 'Question of questions' (Kassakian 1981) and many simulators have evolved after realizing its importance. General purpose simulators such as MATLAB-SIMULINK, SIMNON etc. are widely used in various engineering disciplines while PSPICE, SABER etc. are used especially in electrical circuit analysis. Simulators like EMTDC, EMTP are mostly used in power system area and a few circuit simulators such as ATOSEC, PSIM, SPICE, SIMUPELS, KREAN are dedicated to power electronic converter systems.

2.1 *Micro- and macro-modelling*

All currently available power electronic simulators using either software packages such as EMTP, PSPICE, SIMNON and MATLAB-SIMULINK or dedicated power electronic simulators such as ATOSEC5, and PSIM use a macro-modelling approach for modelling power semiconductor elements . The micro-model used for the semiconductor switch can be any one of the following classes.

  (i) Ideal;

 (ii) two-valued resistor;

(iii) two-valued inductor;

(iv) subcircuits;

 (v) equations describing switch.

The ideal model gives a simple idea of short circuit when the switch is 'on' and open circuit when the switch is 'off' (Lai 1994). Two-valued resistor (Kang & Lavers 1994) and two-valued inductor (Rajagopalan 1988) models implement the low impedance in the 'on'-state and high impedance in the 'off'-state. A number of sources, resistors, and capacitors make the switch model with subcircuits, while the nonlinear equation, representing the switch is used in the equation switch model. This modelling approach has been used and validated by several researchers for the analysis of large power electronic converters. It is useful to develop a library of models for the various components of the power electronic system, namely, power semiconductor devices, electrical machines, power electronic converters and power electronic control strategies. Such a library of models must be available in a user-friendly environment which can be advantageously used for building power electronic converter circuits and systems for analysis. These user-friendly models are now available for various simulators such as SABER, PSIM and SIMUPELS etc.

2.2   *Time/frequency domain*

Simulation techniques can be based on the time domain or the frequency domain principle. To study the switching operation effect on power converters or high switching frequency operated converters time domain analysis is usually used (Liu *et al* 1993; Tymerski 1993). Analysis, design, steady state or transient state conditions of the controller or power electronic converter are usually studied by frequency domain based simulation techniques (Maranesi 1992). It is not always easy to derive models for frequency domain based analysis for complex power converter topologies.

2.3   *Circuit analysis approach*

Simulation techniques are based on any one or two of the following circuit analysis approaches.

2.3a   *Circuit modelling techniques:*   (i) State variable approach (SVA); (ii) nodal analysis (NA); (iii) modified nodal analysis (MNA); (iv) state-space averaging technique (SAT).

The SVA has been widely used in the power electronics area. This approach formulates a set of circuit equations in terms of branch voltages and currents and they are called state equations. These state equations can be further solved by any numerical method. SVA approach has been used by dedicated power electronic simulators such as ATOSEC5 (Rajagopalan & Sankara Rao 1975), SCRIPT (Oms *et al* 1989), COSMIR (Hsiao *et al* 1989), CAP (Liu *et al* 1993).

The traditional nodal analysis based approach is used in EMTP in the area of power systems. Nodal analysis based power electronics simulator is also reported as PECAN (Lavers *et al* 1990). However nodal analysis-based simulators are not so efficient as they are incapable of handling current depending elements, performing fast steady state analysis, and it is difficult to obtain the branch currents in many power electronic circuits. These difficulties are overcome to some extent in Modified Nodal Analysis (MNA) as it includes circuit components such as voltage sources and other current depending elements. This approach has been used in many simulators such as PSPICE, SABER, KREAN etc. MNA has been implemented for switch modelling also (Sudha *et al* 1993).

State-space Averaging Technique (SAT) is mainly based on state averaging concept. This technique converts a nonlinear, time varying, switched circuit into a equivalent circuit by averaging state variables. It is widely used for the analysis of switched mode converters (Middlebrook 1989). Direct implementation of SAT is usually done in SPICE simulator. Recently, computer-aided modelling technique of *dc-dc* converters, using SAT, is also proposed (Chapman & Cichards 1992).

2.3b   *Symbolic math analysis (SMA):*   In Symbolic Math Analysis, the most common modes of operation of a converter topology are identified together with the corresponding equations for the mode changes. The analysis consists in solving the system equations from an initial operating mode with arbitrarily chosen or known initial conditions through a pre-determined sequence of the different operating modes for a whole cycle of operation.

The analytical solutions for each mode of operation is derived for each mode in terms of the unknown initial conditions. Each mode of the system is governed by a set of total differential equations. Generally, the conditions for mode changes are governed by nonlinear equations. Therefore, the solution for a complete cycle requires the computation of all the initial conditions through an iterative procedure. With the availability of general purpose software such as MATLAB and its toolbox called Symbolic Math Analysis or others such as Mathematica, Mapel etc., it is now possible to use the sequential method of analysis for the generation of design curves for new circuit topologies (Rajagopalan 1995). This approach has been used to develop the switch power converter simulator named SPANTRAN (Liberatore *et al* 1991).

There are a few other analytical methods which are used for power electronic converter simulation. Sample-data modelling techniques have been specifically used for the study of *dc-dc* converters which permit derivation of suitable transfer functions of the system thereby permitting traditional control system design methods to be used for the tuning of control parameters (Verghese *et al* 1986).

2.4 *Numerical methods*

Various numerical integration methods have been used in power electronic simulators. Traditional methods such as Newton–Raphson, Runge–Kutta, Backward Differential Formula (BDF), Backward Euler, Trapezoidal, Implicit higher order BDF named Gear etc., are implemented in simulators (Mohan *et al* 1994).

3. Simulators

Many simulators are now available for the simulation of power converters. In the following section, some of the simulators with which authors are familiar are described.

3.1 *General purpose simulators*

3.1a *MATLAB-SIMULINK:* MATrix LABoratory (MATLAB) is a high performance iterative package for scientific engineering numerical computation (Matlab 1992). MATLAB integrates numerical analysis, matrix computation, signal processing and graphics in an user-friendly environment. SIMUlation LINK (SIMULINK) within the MATLAB works in two phases, (1) model definition (2) model analysis (Simulink 1992). It constructs simulation by building blocks or modules. SIMULINK has graphical interfaces for building blocks. It uses the simplified switch models for power electronic converters. It uses the state equation models to solve any continuous/discontinuous, linear/nonlinear system. This simulator has a great facility to exchange the data between MATLAB and SIMULINK. It generates the graphical output using built-in functions. Simulink simulation slows down for the large power electronic converter systems because of its constant passing of data to and fro from MATLAB. The user-friendly graphical capability of MATLAB and various controller design functions are of great advantage in this simulation environment.

3.1b *SIMNON:*   This simulator provides the time domain responses. It requires a separate tool for dealing with frequency domain analysis. It constructs simulation by building blocks or modules. It uses the state equation models to solve any continuous/discontinuous, linear/nonlinear system. It generates the graphical output using built in functions. The simplified switch models are used in power electronic converter's simulation. SEMAS is an example of a software that has been built in SIMNON environment for the time domain analysis of power electronic systems (Ba-Razzouk *et al* 1993). New SIMNON for windows further provides user-friendly interfaces (Simnon *et al* 1993).

3.1c *EMTP:*   Electro Magnetic Transients in Power system (EMTP) is widely used in large scale power systems (Meyer & Liu *et al* 1974) . This simulator has models of transformer, transmission lines, machines and switch models such as diodes, thyristors etc. The nodal approach is used for solving the circuit equations. Here, trapezoidal or factorization methods are used for solving these equations and the time step for integration is specified by the user.

3.1d *EMTDC:*   EMTDC is basically a circuit analysis program (EMTDC 1986). It can be used to model any dynamic systems that can be constructed from building blocks consisting of integrators, limits, logic functions, delays etc. known as Continuous System Modelling Functions (CSMF). Furthermore, electric circuits and CSMF can be interfaced to create a model of both the circuit and controls. EMTDC can be used as a time domain simulation program for electric power and control systems. It also supplies methods for frequency scanning. Usually it is used for studying dynamics of power systems with various converters such as HVDC, STATCOM etc.

3.1e *SPICE:*   Simulation Program with Integrated Circuit Emphasis (SPICE) is used widely in electronic circuit analysis (Nagel 1975). It offers the various models, subcircuit libraries, graphical pre-processors and post-processors etc. Both time-domain and frequency-domain analysis are possible in SPICE. The circuit equations are developed by nodal approach and Gaussian elimination technique is used for solving these equations. Voltage dependent switches are modelled by nonlinear resistors. Nonlinearities can be resolved within the program. However the choice of Trapezoidal or Gear methods is also possible. Here, the time steps are automatically adjusted within the programs. SPICE is used in power electronics for simulation at low power levels e.g. Switch Mode Supplies. SPICE is available in Personal Computer environment, known as PSPICE. It is very popular for circuit analysis because of its availability at very reasonable cost for classroom environments.

3.1f *SABER:*   This simulator is well suited for nonlinear analog and digital circuits, especially for mixed mode circuits in power electronics (Xu 1990). It offers device models, machine models, models for mechanical parts such as gears, dampers, fans etc. This simulator allows the developing of new models in dedicated description language MAST (Franz *et al* 1990). Designers can use other languages such as , FORTRAN, C, Pascal etc.

for developing the new models. This additional facility of external modelling makes the SABER more useful in power electronics.

3.2 *Dedicated power electronic simulators*

3.2a *ATOSEC5:* Analyse Topolologique des Systemes Electroniques en Commutation (ATOSEC5, French acronym) is power electronics dedicated simulator (Rajagopalan & Sankara Rao 1975). The power electronics system is described in the form of a set of state equations. These state equations are solved by Backward Euler method. In order to ensure the numerical stability the Newton Raphson iterative method is used for the solution. A constant time step is chosen for the solution of state equations. Several subroutines are used for FIRING, BLOCK, EXTFIR for well defining the power electronic systems. This offers a library of semiconductor switches such as diode, transistor, FET, GTO, MOSFET and special purpose switches such as PWM, hysteresis etc. User can define special kind of switch which is not covered in ATOSEC through EXTFIR and it can be used in principle subroutine for the solution. ATSECG interface module can be used to prepare the schematic diagram of the circuit. The circuit element values can be shown in a specific manner and can be used in a data file which can be obtained from ATOSECG interface module. ATOPLOT interface module shows the simulated curves which can be visualized during simulation. Any power electronic system, e.g. circuit with 100 branches, 60 nodes and any 40 semiconductor switches, can be easily handled by ATOSEC. All the programs are written in FORTRAN 77.

3.2b *PSIM:* Power electronic circuit SIMulator (PSIM) is a dedicated power electronic simulator (Jin 1995, 1996). It uses the nodal and trapezoidal approaches for circuit equations. It contains the library of several switches such as diodes, thyristors, transistors and bi-directional switches. It also contains the commonly used circuits such as three phase transformers, thyristor converters, PWM voltage/current source inverters. Apart from using the actual control circuit, transfer function blocks and other function blocks can also be used in PSIM. All programs are written in FORTRAN. The first step is to create or edit the power electronic circuit in OrCAD or SIMCAD and then the circuit is simulated in PSIM and results can be viewed in SIMVIEW. Both SIMCAD and SIMVIEW are written in Microsoft Visual C^{++}.

3.2c *SIMUPELS:* SIMUlation of Power Electronic Systems (SIMUPELS) is a user-friendly simulator built in popular MATLAB-SIMULINK environment (Ba-Razzouk *et al* 1995). Power semiconductor devices are modelled using SIMULINK. It also explains how to use the generated device models to construct any power electronic converter. Nodal approach and mesh type formulation of circuit equations are used. Extensive libraries of converters, control circuits, and electrical machine models are included in the simulator. The simulator contains several examples of simulation of many power electronic systems; in particular a number of case studies dealing with *ac* and *dc* drives are available. All the

3.2d *KREAN:* This simulator is dedicated to power electronics especially for power converters with high switching operation (KREAN 1990; Nilssen & Mo 1990). All switches are modelled as ideal switches. Nonlinear components are described in FORTRAN sub-routine and linked to the main program. Nodal approach is used for developing circuit equations. Runga–Kutta, Newton–Raphson, and sparse matrix methods are used for solving these equations.

3.3 *Power electronic tutor (CATPELS)*

Computer-Aided Tutor in Power Electronic Systems (CATPELS) is a useful multimedia tool which helps the user to understand a simulation and a subject from basic to advanced level. This involves sophisticated computer-aided analysis and design tools. This tutor outlines the major requirements and concepts of a computer-aided learning aid for power electronic systems and gives the characteristics of a prototype of an expert system named CATPELS (Rajagopalan *et al* 1995). The open architecture of the run-time version of the CATPELS software provides not only the basic information for beginners and all the required tools for specialists but also permits a user (an instructor or a researcher) to expand its capabilities and to customize as per specific requirements.

4. Demonstrative examples

A commonly used example of three phase self commutated converter, shown in figure 2a, has been analysed and simulated. Figure 2b shows the simulated model in MATLAB-SIMULINK within SIMUPELS (Ba-Razzouk *et al* 1995) . Here, the power converter models are developed on the basis of switching function theory. The voltages at the converter input are synthesized from the *dc* output as given in (1)–(3) below.

$$e_1 = \frac{Vout}{3}(2.S1\text{-}S3\text{-}S5), \tag{1}$$

$$e_2 = \frac{Vout}{3}(2.S3\text{-}S1\text{-}S5), \tag{2}$$

$$e_3 = \frac{Vout}{3}(2.S5\text{-}S1\text{-}S3). \tag{3}$$

The input current depends on the output currents,

$$iout = i_1.S1 + i_2.S3 + i_3.S5 \tag{4}$$

The appropriate space-vector PWM algorithm for the converter is implemented with the MATLAB files. The other auxiliary components referring to the generation of the space vector polar coordinates, the command circuit avoiding the short-circuit and the clock generator have been grouped in the desired parameters of the PWM generator (pulse frequency, number of pulses and modulation index etc.) in a user-friendly manner. Together with the modules for power converters and PWM techniques some special modules are designed for *ac* filters and loads allowing an easy adjustment of desired component values.
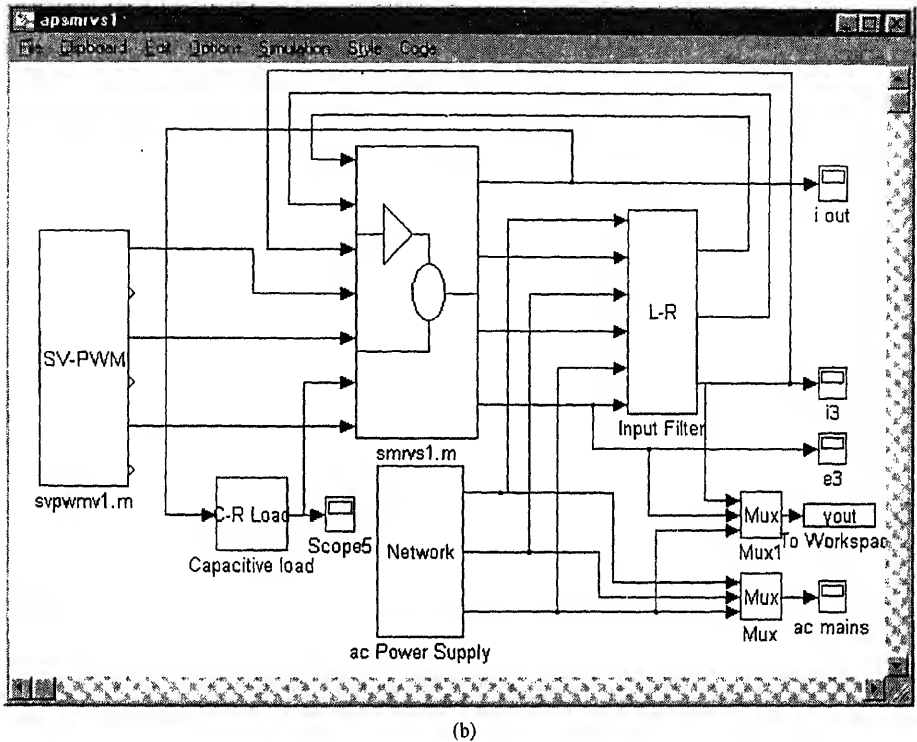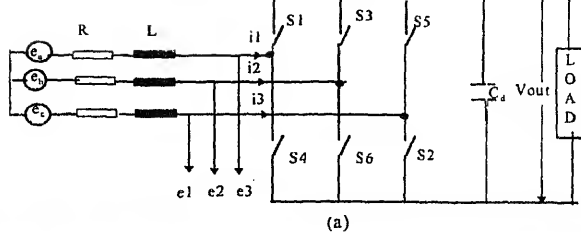
Figure 2. (a) Three phase voltage source converter. (b) Simulated in MATLAB-SIMULINK.

Three more demonstrative examples are provided in figures 3–5. Figure 3 shows the single-phase diode boost rectifier with single-switching device on *dc* side in PSIM. Figures 4 and 5 show the three-phase boost rectifier with single-switching device on *dc* side and simulated in MATLAB-SIMULINK and SABER respectively (Dawande & Dubey 1996). Figure 6 shows a typical case study in CATPELS with various menu options available to the user for *ac–dc* converters.

5. Difficulties and challenges

When the simulation of power electronic converters is carried out, a number of problems may occur due to imperfect models used for devices resulting in lower accuracy. To cite an example, when using macro-models for semiconductor power switches, it is not possi-
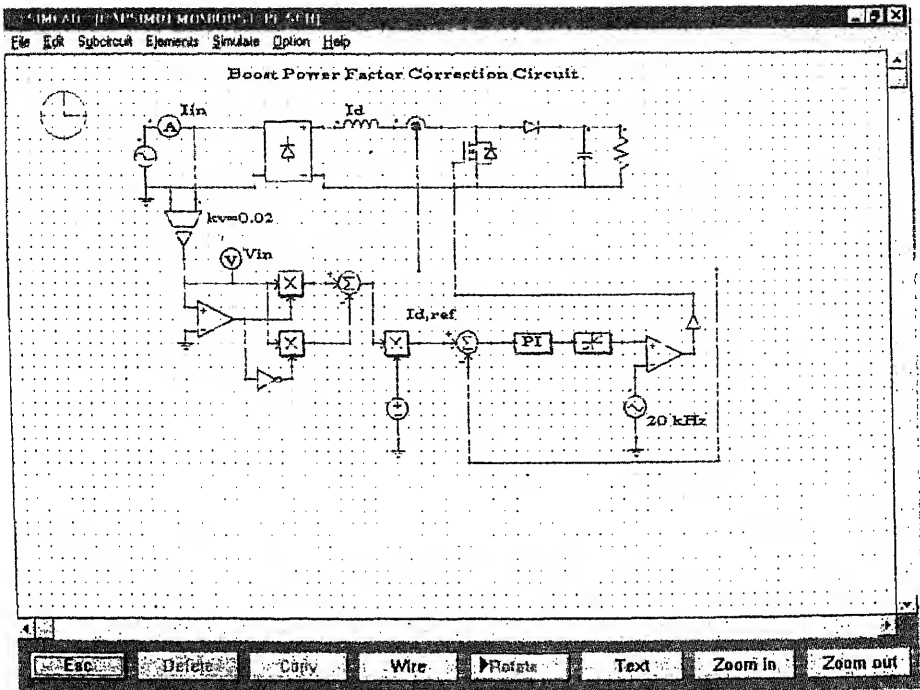
Figure 3.   Single phase boost rectifier in PSIM.
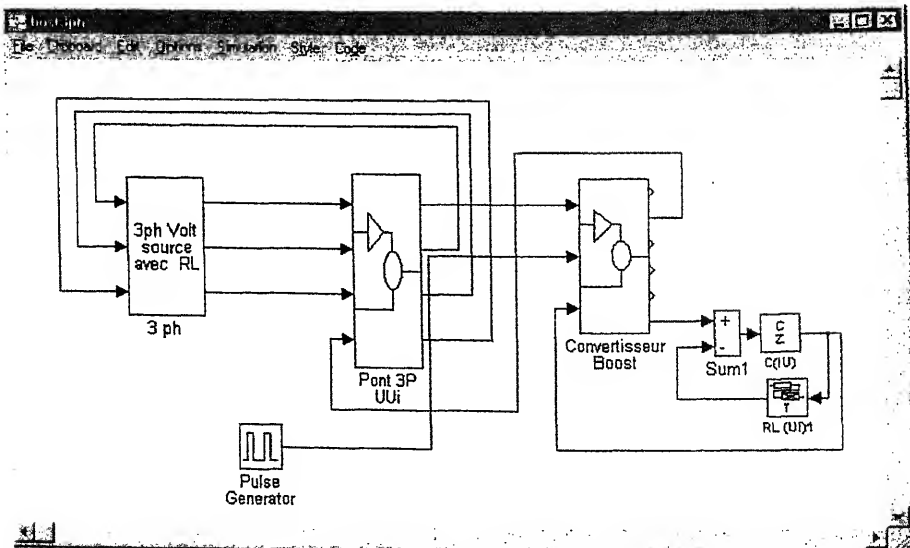
Figure 4.   Three-phase boost rectifier in MATLAB-SIMULINK.
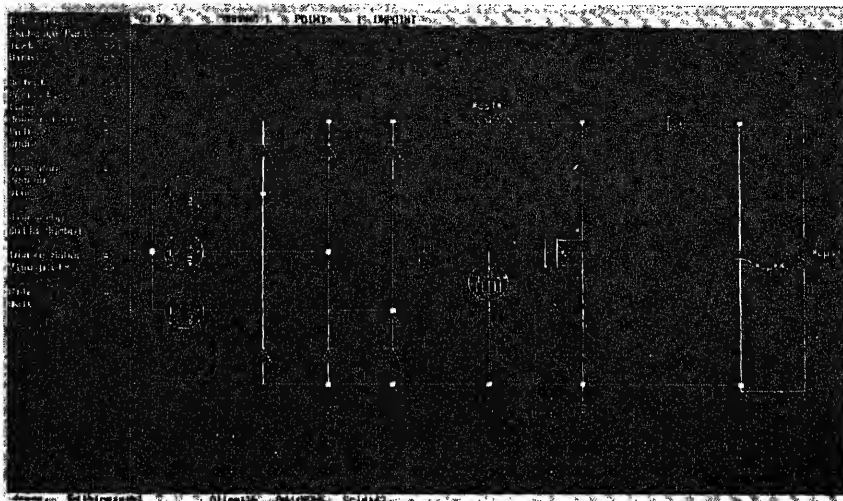
Figure 5. Three-phase boost rectifier in SABER.
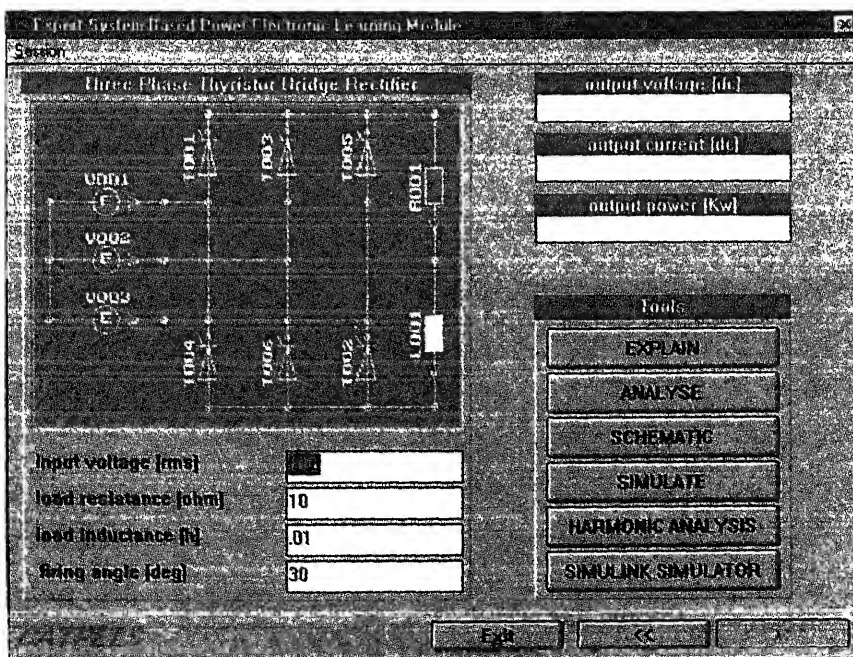
Figure 6. *ac-dc* converter study in CATPELS.

ble to calculate accurately the power losses in the devices during turn-on and turn-off of the devices. In order to calculate accurately these transitions, sophisticated physics-based semiconductor models (micro-models) are required for the switches. Use of such models for the study of power electronic converters increases the complexity, reduces the computational efficiency and even causes numerical unstability. Modelling of bipolar devices such as IGBT, GTO etc. are devoid of carrier diffusion property. Most of the switch models use a single variable equation and do not use the time-dependent partial differential equation for the proper characteristic of the device. Simulation in power electronics becomes difficult when there are sharp discontinuities in the circuits. The study of converter-fed machine requires extensive time for data preparation and solution.

The availability of various simulators dedicated to specific domains have some advantages and disadvantages when chosen for a given application. A universal simulator with a facility to link all the simulators in a user-friendly environment will be welcome. This can provide the study of a particular condition of a power electronic converter system by using a simulator that is most suitable from a library or bank of standard simulators. For the sake of discussion, consider the analysis of a power electronic converter system. The expected standard bank contains several simulators $X_1, X_2, X_3, \ldots X_n$. Suppose X_1 deals with the modelling of devices, which can model the converter with thyristors, MOSEFTs, IGBTs etc., e.g. NIST simulator; X_2 deals with steady/transient state response in time and frequency domain e.g., SIMUPELS; X_3 deals with high voltage application software like EMTP/EMTDC etc. Then the study of any converter system at all levels of application could be easily possible, without any limitations of a certain simulator, in a user-friendly environment. It will be beneficial for all university researchers and industry designers to have such a universal tool readily available for the analysis of any power electronic converter system. However, this approach is feasible only if such simulators are available at reasonable cost with time-sharing possibility on the Internet.

6. Future trends and developments

In the next few years, computer-aided design of power electronic converters will be more needed to reduce the design and manufacturing cost. The system which can be designed to control, implement and test the high speed control algorithms of power electronic converters and which can bring the hardware in the loop simulations in real experiments will be essential for future applications.

Presently, the link of simulation and hardware has been brought out by using dSPACE system in our research group. This system consists of Digital Signal Processor (DSP TMS320C40), input/output boards, graphical tool TRACE etc. The following systematic process is adopted.

A simulated system of harmonic compensation of three-phase converter is developed in MATLAB-SIMULINK. The necessary controller is designed in SIMULINK and grouped in the block of 'compensator discrete' (Yao 1997). After achieving the satisfactory simulated results, the controller is modelled in dSPACE environment with necessary input/output channels which are shown in figure 7a. This has to be built from the selected hardware library of dSPACE. Channels 1-3 sense the three input currents of a three-phase converter,

while channels 3-6 sense the input voltages of a three-phase converter, through ADC board (DS2003). The processed output signals of the controller are delivered to hardware through DAC board (DS2103). The dSPACE Real-Time Interface (RTI) contains all the interface software necessary to integrate the combined development of the MATLAB, SIMULINK and Real-Time Workshop. Code generation from SIMULINK block diagrams is fully automatic. Figure 7b shows a real-time interface which integrates the simulation and hardware. Figure 7c shows a controller model which has been downloaded to DSP. All the necessary input/output signals along with the controller signals are seen through graphical tool called TRACE in dSPACE. The option menu of TRACE for various signals is seen in figure 7c. The implementation of the simulated controller, directly to hardware using dSPACE is successfully done. The actual experimental monitored signals are seen in figure 7d.
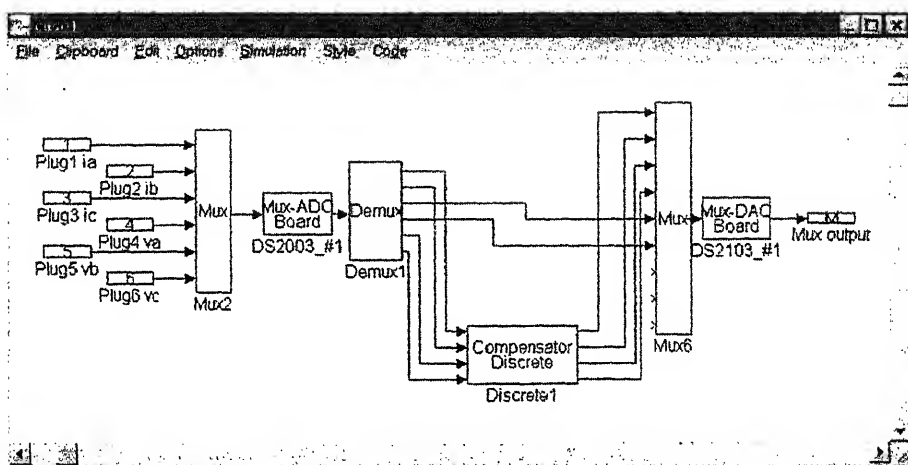
(a)
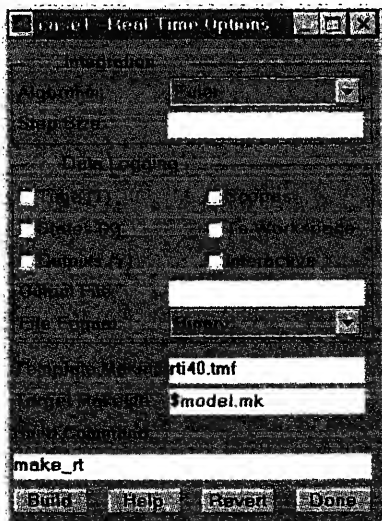
(b)
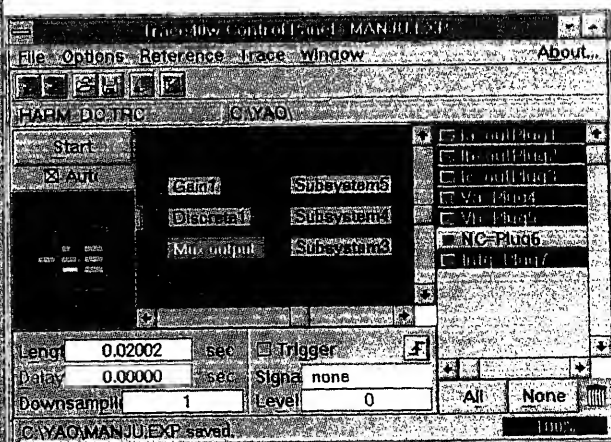
Figure 7. (Continued on next page.)
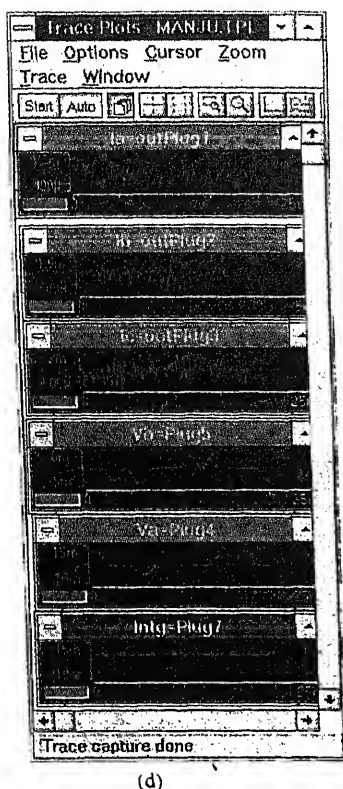
Figure 7. (a) Harmonic compensator model developed in dSPACE. (b) Real Time Interface integrating simulation and hardware. (c) Graphical TRACE tool showing simulated and experimental signals. (d) Monitored experimental results.

7. Conclusions

This paper gives a comprehensive state-of-the-art review of power electronic simulation techniques. It reviews various methods of modeling, circuit analysis approaches and numerical methods used in different simulation techniques. Several general purpose simulators such as MATLAB-SIMULINK, SIMNON, EMTP, EMTDC, SPICE, SABER etc. and dedicated power electronic circuit simulators like ATOSEC5, PSIM, SIMUPELS, KREAN are discussed in brief. Several demonstrative examples using different simulators are provided to show the importance and ease of simulation in power electronic converters. Considering the use of hardware in simulation which may be the essential future demand in power electronic converter systems, some new developments are introduced to show the wide scope of research and developments.

References

Ba-Razzouk A, Debebe K, Rafesthain T, Rajagopalan V 1993 SEMAS: A new simulation module for ac drive systems. *Can. J. Electon. Comput. Eng.* 18: 181–189

Ba-Razzouk A, Debebe K, Neacsu D O, Gheorghe M, Pittet A, Yao Z, Rajagopalan V 1995 SIMUPELS: Simulation of power electronic systems using MATLAB-SIMULINK software. Research Report, Chaire de Recherche Industrielle HQ-CRSNG, UQTR, Trois-Rivieres (Quebec), Canada

Chapman M, Cichards R 1992 A method for improving the speed of simulation of switched mode power supply circuits. *IEE Proc. Colloq. on CAD of Power Electronic Circuits*, London

Dawande M, Dubey G K 1996 Programmable input power factor correction method for switch mode rectifiers. *IEEE Trans. Power Electron.* 11: 585–591

EMTDC 1986 Users' Manual. Manitoba HVDC Research Center, Canada

Franz G A, Ludwig G W, Steigerwald R L 1990 Modeling and simulation of distributed power systems. *IEEE Power Electron. Syst. Conf. 90* pp 606–610

Hsiao C J, Ridley R B, Lee F C 1989 The simulation of switching converters using the new version COSMIR program. *Proc. VPEC Seminar* pp 121–130

Jin H 1995a PSIM: Power electronic system simulator. Software 1995, Department of Electrical and Computer Engineering, Concordia University, Montreal

Jin H 1995b Computer simulation of power electronic circuits and systems using PSIM. *International Conf. on Power Electron.* Seoul, pp 79–84

Jin H 1996 A new computer-aided design tool for switchmode power supplies. *High-Frequency Power Conversion Conf.* Las Vegas, Nevada, pp 117–122

Kang Y, Lavers J D 1994 Power electronics simulation: Current progress and future development. *IEEE PELS 4th Workshop on Computers in Power Electronics* pp 169–174

Kassakian J G 1981 Simulating power electronic systems – An overview. In *Proc. US – Japan Cooperative Science Seminar on Analysis and Design in Power Electronics* (Kobe, Japan) pp 45–54

KREAN 1990 The Norwegian Institute of Technology, Group of Power Electronic and Electrical Machines, N-7034 Trondheim, Norway

Lai J S 1994 Power electronics modeling and simulation. *IEEE 4th Workshop on Computers in Power Electronics*, Invited Paper, 0-7803-2091-3/94, pp 45–55

Lavers J D, Jin H, Cheung R W 1990 Analysis of power electronic circuits with feedback control: A general approach. *Inst. Elec. Eng. Proc.* B137: 4

Liberatore A, Manetti S, Piccirilli M C, Reatri A 1991 A new symbolic approach to the analysis of switch power converters. *European Power Electronics Conf.-91* 4: 489–490

Liu C C, Hsieh J, Chang G H K, Bocek J M, Hsiao Y T 1993 A fast decoupled method for time domain simulation of power converter. *IEEE Trans. Power Electron.* 1: 37–45

Maranesi P 1992 Small signal circuit modeling in the frequency domain by computer aided time domain simulation. *IEEE Trans. Power Electron.* 7: 83–88

Matlab 1992 MATLAB User Guide. The Math Works

Middlebrook R D 1989 Modeling current-programmed buck and boost regulators. *IEEE Trans. Power Electron.* 4: 36–52

Mohan N, Robbins W P, Undeland T M, Nilssen R, Mo O 1994 Simulation of power electronic and motion control systems – An overview. *Proc. IEEE Conf.* 1: 1287–1302

Nagel L W 1975 SPICE2: A computer program to simulate semiconductor circuits. Memo. ERL-M520, Univ. of Berkeley, CA

Meyer W S, Liu T H 1974 EMTP Rule Book. Bonneville Power Administration

Nilssen R, Mo O 1990 KREAN, a new simulation program for power electronics. *IEEE Power Electron. Syst. Conf. Rec. 90* pp 506–511

Nilssen R 1991 Programmable module in simulation programs for power electronic circuits. EPE-91, Firence, Italy, 4: 373–377

Oms F *et al* 1989 SCRIPT simulator from converter to semiconductor. *EPE'89* Aachen, vol 1, pp 207–212

Rajagopalan V 1988 Computer-aided analysis of power electronic systems. *IECON'88*, Singapore, Invited Paper, 88CH2602-1, pp 528–533

Rajagopalan V 1987 *Computer-aided analysis of power electronic systems.* (New York: Marcel Dekker)

Rajagopalan V 1995 Modeling and simulation of power electronic converters for power supplies. *IEEE Conf. Rec. IECON 95* Invited Paper, pp 27–32

Rajagopalan V, Sankara Rao K 1975 *ATOSEC Users' Manual*, Département de génie électrique, Université du Québec à Trois-Rivières, Canada

Rajagopalan V *et al* 1990 User friendly dedicated power electronic converter simulator. *IEEE Workshop on Computers in Power Electronics*, pp 183–204

Rajagopalan V, Debebe K, Chandrasekaran A, Sudha S A 1992 User-friendly dedicated power electronic converter simulator. *IEEE Trans. Ind. Electron.* 39: 55–62

Rajagopalan V, Yao Z, Brillon D, Doumbia M L 1995 CATPELS: Computer-Aided Tutor for Power Electronic Systems. Université du Québec à Trois-Rivières, Trois-Rivières, (Québec), G9A 5H7, Canada

Simmon 1993 Simnon for Windows. SSPA Systems, Sweden

Simulink 1992 SIMULINK User Guide. The Math Works

Sudha S A, Chandrasekaran A, Rajagopalan V 1993 New approach to switch modeling in the analysis of power electronic systems. *Inst. Elec. Eng. Proc.* B140: 115–123

Tymerski R 1993 A fast time-domain simulator for power electronic systems. *IEEE APEC-93*, pp 477–483

Verghese G C, Elbuluk M C, Kassakian J G 1986 A general approach to sample-data modeling for power electronic circuits. *IEEE Trans. Power Electron.* 1: 76–89

Xu C 1990 Modeling and simulation of power electronics using SABER. Presented at First European Meeting of SABER User's Group, Newbury, England

Yao Z, Dawande M, Rajagopalan V 1997 Controller design for advanced reactive power compensators based on input-output linearization. Acceptance received for *IEEE Power Electron. Syst. Conf. 97* pp 936–941

Recent advances in var compensators

GÉZA JOÓS

Department of Electrical and Computer Engineering, Concordia University, 1455 de Maisonneuve W, Montreal, Canada H3G 1M8
e-mail: geza@ece.concordia.ca

Abstract. Static var compensators have been, for many years, an essential component in the operation of power transmission systems. They are part of a family of devices known as Flexible AC Transmission System (FACTS) devices. The advent of large capacity force-commutated semiconductor switches allows many developments in power electronic converters to be applied to the implementation of high power compensators. This paper describes the principles of controlled reactive power compensation, particularly in the context of power systems. It focuses on active static power converter-based compensators and discusses issues related to the power circuit topology and control techniques, including the impact of Pulse Width Modulation (PWM) techniques. Compensators based on current and voltage source converters and on *ac* controllers, both in the shunt and series configurations, are covered. Methods to enhance power capacity using multi-level and multi-pulse arrangements are discussed.

Keywords. Reactive compensation; power electronic converters; power systems.

1. Introduction

Reactive power (var) compensation has long been recognized as an essential function in the operation of power systems. At the distribution level, it is used to improve the power factor and support the voltage of large industrial loads, such as line commutated-thyristor drives and electric arc furnaces. Reactive power compensation also plays a crucial role at the transmission level in supporting the line voltage and stabilizing the system. Rotating synchronous condensers and mechanically-switched capacitor and inductor banks have been replaced in the 1970s by thyristor-based technologies: in typical installations, a thyristor-controlled reactor (TCR) provides variable lagging vars, and fixed or thyristor-switched capacitors (TSC) provide the leading vars. The combination of both devices in parallel allows continuous control of vars over a wide range of values, from leading to lagging vars (Gyugyi 1979). A large number of units have been successfully installed and operated for many years. At the same time, the potential of var compensators based on

static power converters have also been recognized and a number of configurations proposed and investigated (Gyugyi 1979).

However, thyristor technology only allows the implementation of lagging var generators, unless complex force-commutation circuits are used. This drawback has been eliminated with the introduction of Gate Turn-Off (GTO) thyristors (Larsen *et al* 1992). This has allowed the development of a number of configurations based on the use of synchronous voltage sources (Gyugyi 1993). Prototype GTO-based var compensation units, known as STATCOMs have been installed and tested by utilities (Schauder *et al* 1995). The STATCOM and other static var compensators have recently been grouped, together with other types of transmission system control devices, under the heading of Flexible AC Transmission System (FACTS) devices.

Reactive power compensators are typically connected in shunt across transmission and distribution systems. An alternative connection, the series connection, has recently received much attention from utilities (Gyugyi *et al* 1996). Technological solutions have been developed to solve problems associated with insulating the equipment from ground and the full potential of series connections can now be exploited. The latest development in var compensation technology has been the combination of series and shunt static compensators into one unit, known in the area of power systems under the name of Unified Power Flow Controller, or UPFC (Gyugyi 1992).

Static power converters have been successfully applied to a large number of power conversion problems at low and medium power levels. However adapting these solutions to high power transmission and distribution levels raises special issues. Although the capacity of power semiconductor switching devices has gradually increased, large ratings still require combining devices in series and parallel. In addition to the large power handling capacity, static compensators must have very high efficiency, since losses have a negative impact on both the initial and operating costs of the power system. Switching losses are therefore a primary concern and switching frequencies must therefore be kept low. This may result in large harmonic waveform distortion, unless special power circuit configurations are used.

This paper reviews the various methods available for generating reactive power (var) by means of force-commutated static power converters, taking into account the above constraints. It discusses topologies suitable for use with devices such as GTOs and the more recently available high power IGBTs and addresses switch gating issues, including the use of Pulse Width Modulation (PWM) techniques. Methods for designing high power converters suitable for transmission level compensation are presented, particularly multi-level and multi-module topologies.

2.   Principles of var compensation

Var compensation can be viewed as the injection of reactive power, leading or lagging, into the *ac* system. In its simplest form, reactive power injection is achieved by inserting fixed capacitors or inductors in either series or shunt into the *ac* system. Assuming a compensation reactance X_c is inserted in a transmission system, the generated var Q_c is derived as follows.
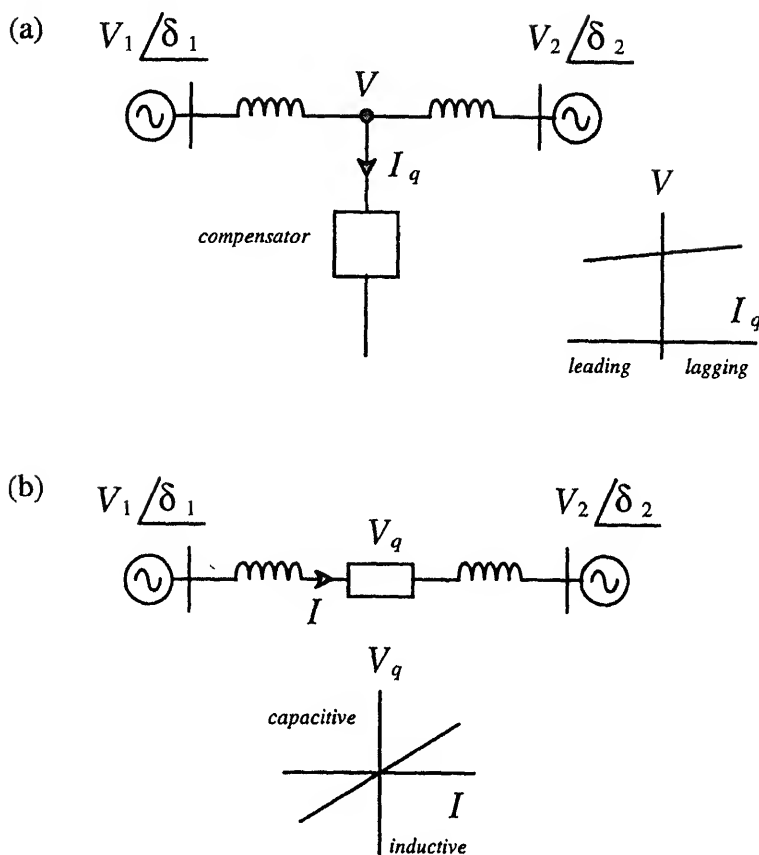
Figure 1. Principles of var compensation in transmission systems. (**a**) Shunt compensation. (**b**) Series compensation.

(a) For the shunt connection, figure 1a, a reactive current I_{cq} is generated, allowing in particular line voltage support at the point of connection, V_c:

$$I_{cq} = V_c/X_c,$$
$$Q_c = V_c^2/X_c.$$

(b) For the series connection, figure 1b the reactive impedance X_c partially compensates the line reactance, and a reactive voltage V_{cq} is inserted in series, the current I_c being the line current:

$$V_{cq} = I_c X_c,$$
$$Q_c = I_c^2 X_c.$$

In addition to the lack of controllability of the reactive power injection, fixed capacitive compensation can lead to ac system instability, such as in the phenomenon known as subsynchronous resonance, or SSR, associated with series compensation (Ooi & Dai 1993;

impedance must be varied. Equivalently, a variable current or variable voltage is injected into the system, emulating a variable reactance, figure 1b.

The apparent reactive impedance of a fixed element can be varied using *ac* switches, or *ac* controllers. On the other hand, the current or voltage required to emulate a variable reactance can be injected into the *ac* system by means of synthetic sources, which can be realized using static power converters (Agaki *et al* 1984; Campos *et al* 1994). In addition to providing reactive power, these active compensators can also supply real power, either transiently or for a number of periods of the *ac* supply. This real power can be used to dampen power system oscillations or temporarily support the power system voltage under fault conditions. Furthermore, since the compensator is fully controllable, resonant frequencies associated with the use of capacitors are eliminated and the potential for instability suppressed.
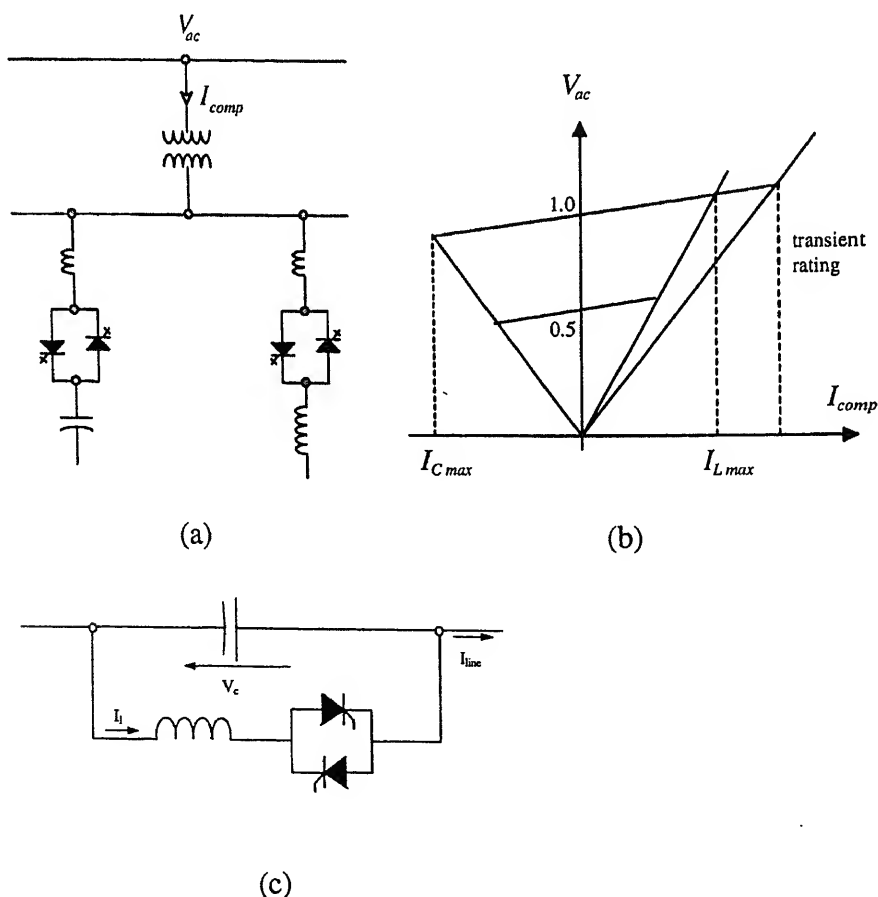
Figure 2. Thyristor-controlled variable reactance compensators. (a) Thyristor-controlled reactor (TCR) and thyristor-switched capacitor (TSC) shunt compensator. (b) Transfer characteristics of a TCR with fixed capacitor. (c) Thyristor-controlled series capacitor compensator (TCSC).

3. AC controller-based structures

3.1 *Conventional thyristor-controlled reactor*

The basic scheme, figure 2a, consists of an *ac* controller which varies the apparent induc-
tance of the inductor as reflected onto the *ac* line (Gyugyi 1979). The thyristor-controlled
reactor (TCR) provides continuously controllable lagging vars and is biased using fixed,
or more often, thyristor-switched capacitors (TSC). The injected vars can therefore be
continuously adjusted from leading to lagging, figure 2b. However, the var injection or
voltage regulation capability of the compensator is limited by the value of the reactance
and is therefore line voltage dependent, figure 2b. Advantages of the system however in-
clude ruggedness, high efficiency, good dynamic performance and competitive cost. The
disadvantages include injection into the line of large low frequency harmonic currents,
dominant being the 5th and 7th components (300 and 420 Hz for a 60 Hz system) for the
basic *ac* controller. Harmonics can be moved to higher frequencies by paralleling units
and using special transformer configurations. Harmonic currents can be reduced by means
of tuned LC filters. These however are costly and can cause voltage oscillations resulting
from the added system resonant frequencies.

3.2 *Force-commutated ac controller structures*

An alternative to the thyristor-based *ac* controllers is the force-commutated *ac* controller,
figure 3a (Jin *et al* 1994; Venkataramanan & Johnson 1997). The use of force-commutated
switching devices removes the requirement for operating the converter in synchronism with
the ac supply and allows gating the switches more than once per cycle. Arbitrary gating
patterns can be implemented, particularly PWM patterns (Jin *et al* 1994). The principles of

(a)   (b)
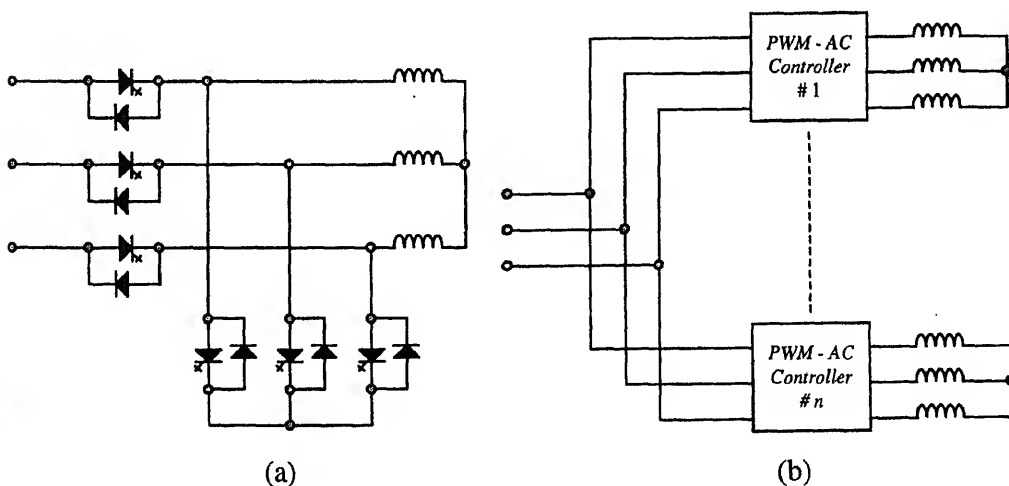
Figure 3. Variable reactance scheme based on force-commutated *ac* controllers (Bias ca-
pacitors are not shown). (**a**) Basic *ac* controller structure. (**b**) Multi-module converter.

PWM pattern generation are explained in § 6.2. Assuming the inductor current is sinusoidal, a pattern with constant duty cycle yields ac line side currents that only contain harmonics around the switching frequency and its multiples. This pattern is simple to implement and allows control of the equivalent inductance; therefore, the amount of vars absorbed can be varied from 0 to a maximum value. Since losses associated with switching large currents at high voltages increase significantly with switching frequency, this frequency must be kept low, typically a few hundred hertz for GTOs. In order to reduce the distortion of injected currents, while keeping the switching frequency low, elementary modules are connected in parallel, figure 3b, and gated so that harmonics are minimized (see discussion in § 7). This also allows the realization of var compensators of large ratings (Lopes et al 1996).

3.3   Thyristor series controlled capacitor

The dual of the TCR is the thyristor-controlled series capacitor (TCSC), figure 2c. It uses a thyristor-based ac controller to continuously adjust the apparent reactance seen from the line. Such units have been successfully installed in transmission systems. However, they have the same limitations as TCRs, including harmonic injection. Performance and harmonic reduction may be enhanced by using force-commutated switches.

   Although ac controller-based compensators are rugged and simple to control, operating regions are determined by the reactances used and by the line conditions, figure 2b. Converter-based structures are more versatile and flexible and are therefore receiving more attention.

4.   Active compensator topologies

The availability of controllable voltage and current sources, realized using static power converters, allows the implementation of static compensators which can emulate the characteristics of a variable reactance, (figure 4). Advantages over the variable reactance structures include the following.

(a) The performance characteristics are improved, particularly reactive power generation Q_{ref} capability, and are independent of the line parameters and line steady state operating voltages and currents, figure 5b.

(b) Energy storage devices, such as battery banks (Walker 1990) and superconducting magnet energy storage systems (SMES) (Zhang & Ooi 1993) can be incorporated into the units. This allows the absorption and generation of real power, P_{ref}, which can be used for power system damping and voltage support under fault conditions.

   In the shunt connection, figure 4a, the power converter system behaves in a manner similar to a synchronous machine operated as a synchronous condenser. If the converter is considered a voltage source behind an inductance, varying the output voltage is equivalent to varying the internal voltage of the synchronous generator, figure 5a. The shunt connection
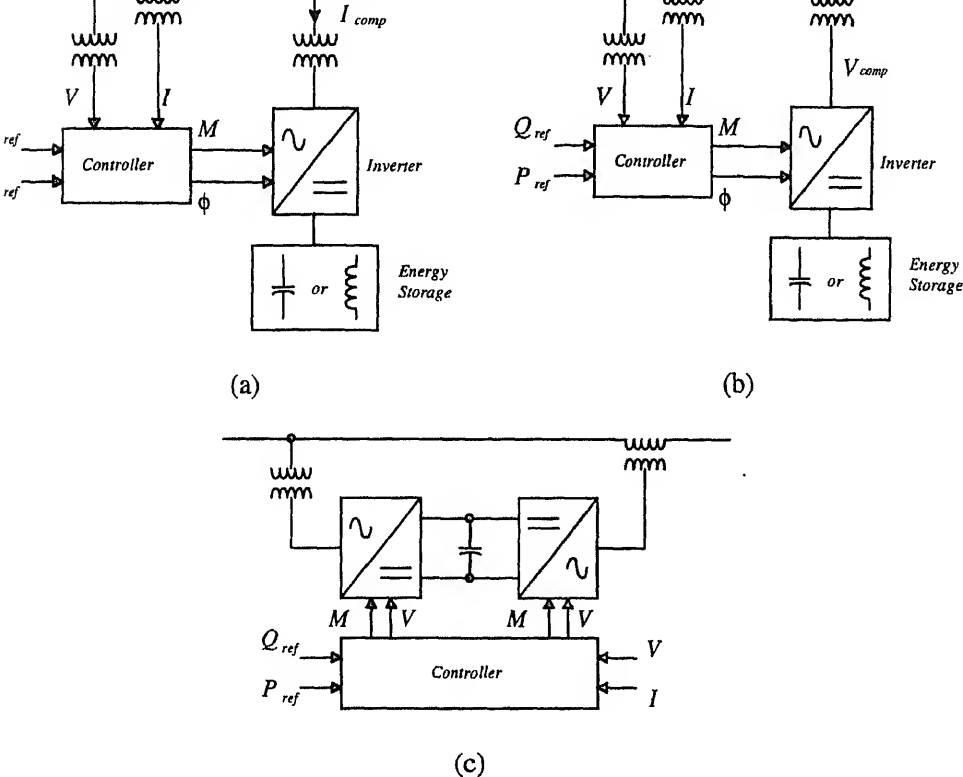
Figure 4. Principles of active compensation based on static power converters. (a) Shunt connection. (b) Series connection. (c) Series-shunt connection (UPFC).

s long been the preferred connection: one end of the compensator can be connected to ound and fault currents do not flow through the compensator.

The series connection, figure 4b, offers a number of advantages over passive series mpensation at a cost that is becoming competitive in demanding applications (Gyugyi al 1996). These include: (i) the elimination of the risk of resonance normally asso- ted with fixed series compensation (SSR); (ii) the possibility of injecting real power o the system and modifying the apparent resistance of the line. This improves the wer transmission system characteristics, including the possibility of increasing the X/R io.

The combination of series and shunt connected converters, connected to a common bus, figure 4c provides the advantages of both the series and the shunt configuration d adds another degree of control. This structure is known as the Unified Power Flow ntroller or UPFC (Gyugyi 1992). In addition to its operation as a var compensator, s scheme can also be used as a phase shifter. Although the *dc* bus can be either of the ltage or current source type, most of the structures that are investigated are based on the ltage source configuration. The voltage source *dc* bus, figure 4c, is the standard config- ation for the UPFC structure. Inverter structures and their characteristics are discussed § 5.
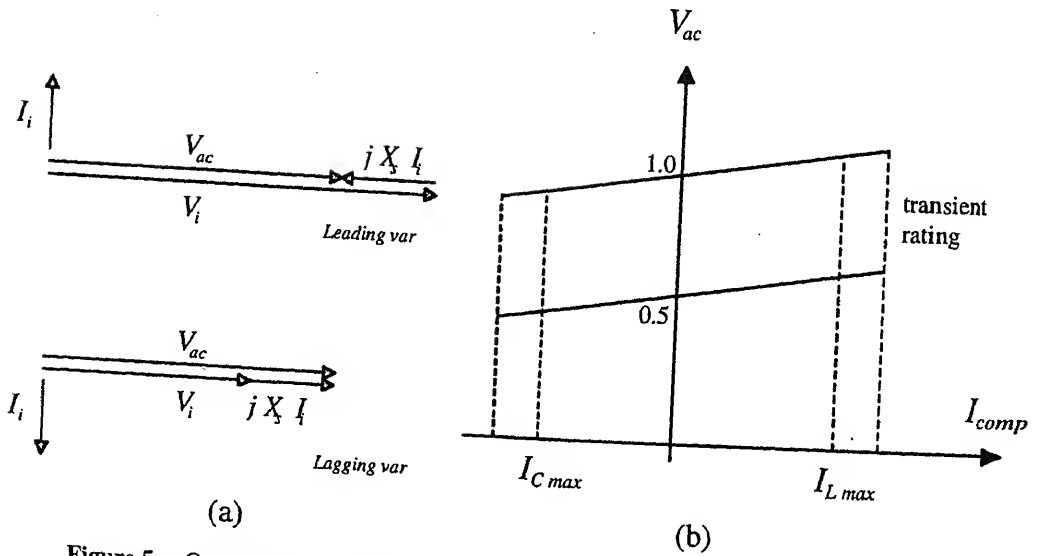
Figure 5. Operation of a voltage source-based compensator. **(a)** Phasor diagram illustrating var generation principles. **(b)** Transfer characteristics.

5. Converter structures

5.1 *Current source type structures*

In a thyristor-based implementation of the compensator, the basic converter structure consists of a 6-pulse ac/dc rectifier/inverter loaded with an inductor. Only controlled reactive power is exchanged between the ac line and the inverter and the compensator draws the real power required to supply inductor and inverter losses from the ac source. The dc bus is of the current source type and the converter structure is similar to that of figure 6a. However, thyristor-based units can only provide lagging reactive power and line side current harmonics are large, figure 7. The compensator has been used in special cases, such as SSR mitigation, for power system damping.

(a)
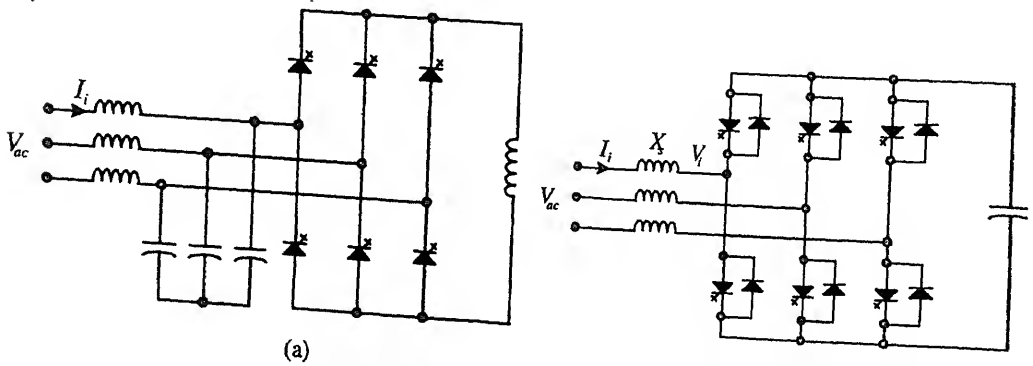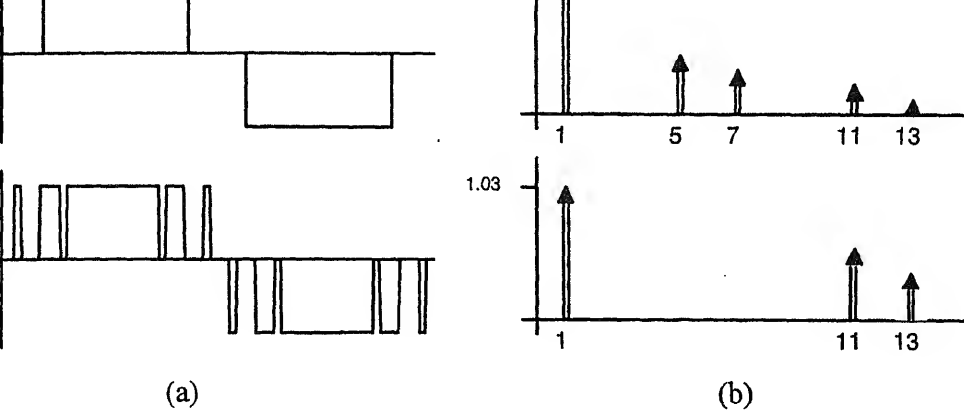
Figure 7. Static power converter output waveforms. AC line voltage for voltage source or *ac* line current for current source. **(a)** Waveforms for single pulse patterns and pulse width modulation (SHE) patterns (5th and 7th harmonic elimination). **(b)** Harmonic spectrum for single pulse patterns and pulse width modulation (SHE) patterns.

The lagging reactive power limitation is removed when thyristors are either force-commutated (Walker 1986), or replaced by force-commutated devices, such as GTOs, figure 6a (Edwards *et al* 1988). The use of such devices also allows the implementation of PWM patterns to (a) control the level of injected current, (b) limit the amount of harmonic currents generated (Moran *et al* 1989a; Espinoza & Joos 1994). However, the structure requires an input capacitor and a line filter reactor, with the associated resonant frequency. Furthermore, in the standard control scheme, the current in the *dc* inductor is maintained constant at the rated value, resulting in significant losses, even when the compensator supplies no reactive power. This has led to the development of compensators based on superconducting magnet energy storage devices, SMES (Zhang & Ooi 1993). This technology however is complex and costly and the preferred compensator configuration is based on voltage source structures.

Voltage source type structures

These structures use a voltage source, a capacitor or a battery bank, on the *dc* side of the *dc* inverter and reflect an *ac* voltage on the line side, figure 6b. Therefore, an *ac* inductor must be used to couple the converter to the line. Part or, in some implementations, all of the inductance is provided by the transformer leakage reactance. As such, the structure is similar to that of a synchronous condenser, and the converter may be referred to as a synchronous voltage source.

As in a synchronous condenser, the current injected into the *ac* supply is controlled by the magnitude of the reflected voltage, figure 5a: a voltage larger than the *ac* supply results in injected vars (leading or capacitive); the converse, or lagging reactive power, results if the voltage is smaller. Methods for controlling this voltage are explained in § 6.3. Since the bus voltage is controlled and independent of the line voltage, the steady state operation

of the compensator, in particular the compensating capability, illustrated in figure 5b, is independent of the line conditions (compare with figure 2b for a TCR-TSC conventional compensator). The compensator therefore provides a better dynamic response and the voltage support in the case of operation of the power system under fault conditions is improved.

The converter operates as a reactive power source, and in steady state does not supply or absorb real power. Losses however must be supplied either: (i) from the *ac* supply for a self-controlled *dc* bus, the preferred solution, in which case an additional *dc* capacitor voltage control loop must be used and a small amount of real power flows into the converter to cover losses (Moran *et al* 1989b); (ii) from a separate *dc* source, usually an *ac/dc* rectifier. If the var compensator is required to supply real power under transient conditions, battery banks can be connected on the *dc* side (Walker 1990).

Advantages of the scheme include low harmonic current injection, if harmonic voltage components are minimized (see §§ 6.2 and 7), and high dynamic performance. Losses are lower than for current source structures. Weaknesses include the risk of *dc* short circuits, when both switches in a leg conduct, and the requirements for output short circuit protection. Current source structures are more rugged: the short circuit current is limited, the *dc* bus being a current source.

6.   Converter control techniques

The basic requirement of the compensator is to generate vars, by controlling either the voltage or the current generated by the converter. Since most of the compensators are based on the voltage source topology, the discussion on control techniques will focus on this topology. However, the conclusions regarding gating patterns are equally applicable to the current source topology: by virtue of duality, current source inverter (CSI) current patterns are identical to the line to line voltage patterns of the voltage source inverter (VSI). However, the gating of a VSI is generally simpler than that of a CSI: this is the result of the bidirectional current capability of the switches in the VSI, which ensures that there always is a return path for the *dc* bus current; therefore, there are no restrictions on the gating pattern. Switches in the CSI however are unidirectional devices and therefore require complementary gating to satisfy the requirements for *dc* bus current continuity.

6.1   *Single pulse gating*

Thyristors in rectifiers and inverters can only be gated once per cycle and line current must always be lagging the voltage to ensure line commutation. If thyristors are replaced by force-commutated switches such as GTOs, this restriction is lifted and both leading and lagging currents can be generated. Gating of each switch in a leg for half the *ac* supply period leads to a single pulse per period: this is the conventional means of controlling large static synchronous condensers, figure 7. An important advantage is that this method leads to the lowest switching losses and the highest efficiency. However, the pattern is fixed and the fundamental or desired component of the voltage (or current in a CSI) depends upon the *dc* bus voltage (or current). Furthermore, the harmonic content of the inverter
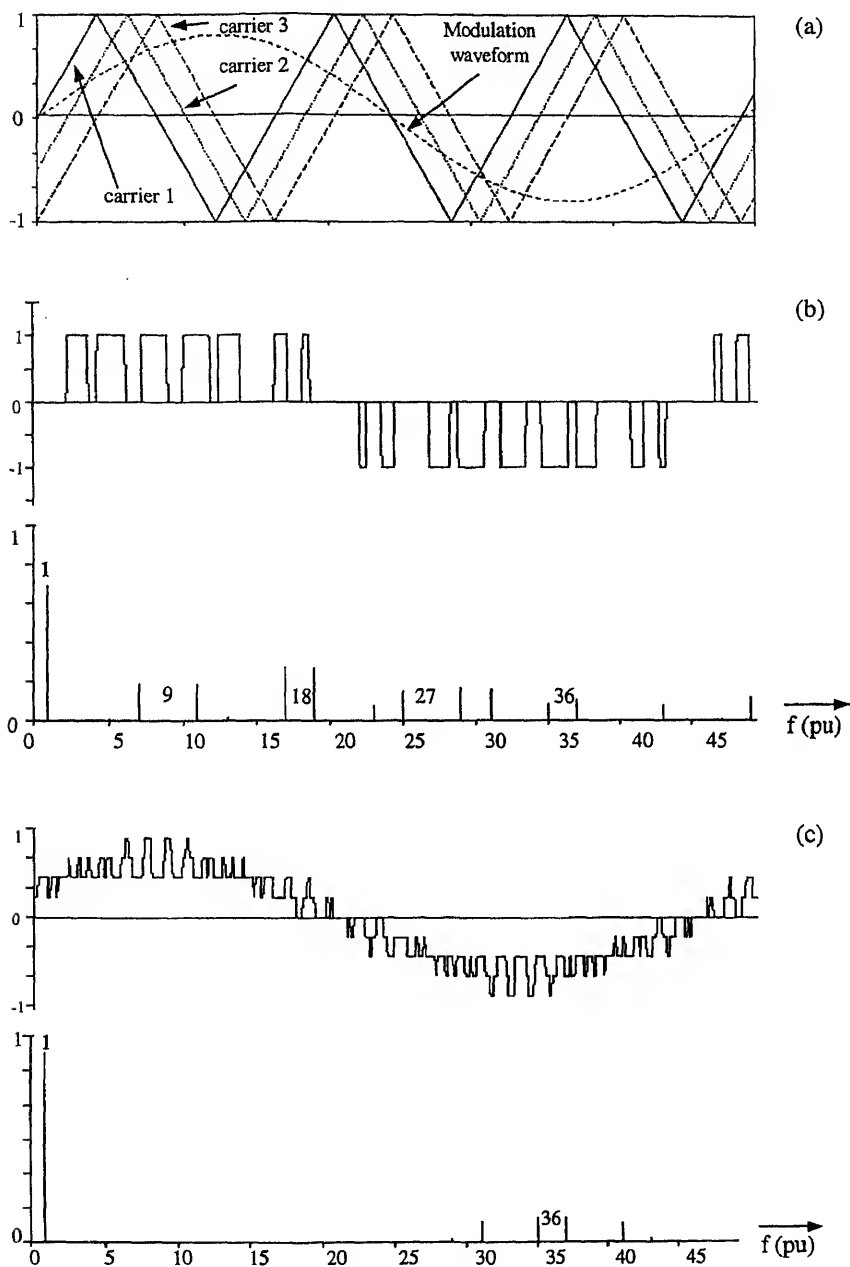
Figure 8. Multi-module pattern generation by carrier phase shifting. (**a**) Principles of Sine Pulse Width Modulation (SPWM) and phase shifted carriers (8 carriers, with only 3 shown). (**b**) Output voltage of individual modules and harmonic spectrum (carrier to fundamental frequency = 9). (**c**) Total output voltage and harmonic spectrum (4 units).

(a)
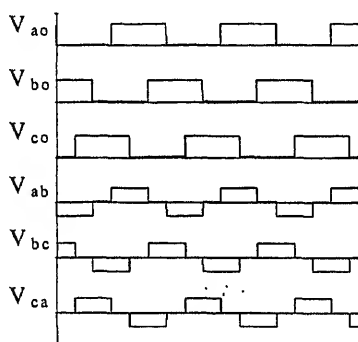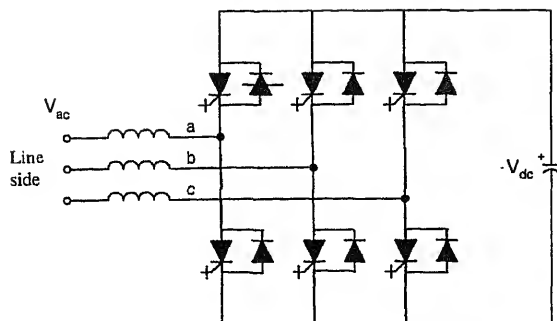
(b)
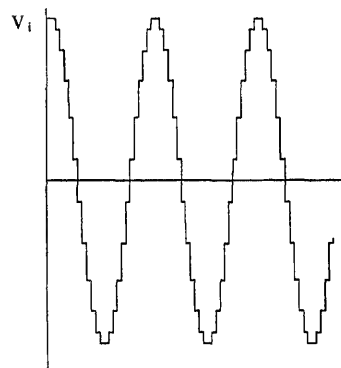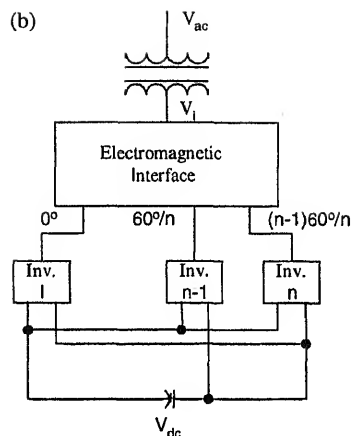
Figure 9. Multi-pulse compensator – voltage source topology. (a) Structure of individual units and output voltage waveforms. (b) Harmonic cancellation by transformer phase shifting and inverter output voltage waveform (24 pulse, $n = 4$, dominant harmonics at the 23rd and 25th).

output voltage is large resulting in large injected harmonic currents. Methods to redu the harmonic content include large filters, a solution that is uneconomical or multi-pul techniques, discussed in § 7.

6.2 *PWM techniques*

In order to vary the fundamental component of the output voltage for a fixed *dc* bus voltag Pulse Width Modulation (PWM) is used. A number of techniques have been develope among which: (i) selective harmonic elimination (SHE), (figure 7) (Kojori *et al* 199 (ii) carrier PWM, the most common being the sine PWM (SPWM), (figure 8a) (Kuang Ooi 1994; Suzuki *et al* 1997), and (iii) space vector modulation (SVM) (Bakhshai *et* 1997). For high power applications, SHE has been successfully applied, since it minimiz switching frequency for a given harmonic content. Criteria for selecting an appropri

(a) *Instantaneous on-line control:* SPWM and SVM allow instantaneous control of the inverter output voltage; patterns generated by a SHE algorithm are generally stored in an EPROM or equivalent, which leads to a slower dynamic response.

(b) *Harmonic minimization:* Direct, selective and precise harmonic elimination can only be obtained with the SHE techniques; other patterns rely on the fact that the dominant harmonics are related to the switching frequency and its multiples. The switching frequency is equal to the carrier frequency for SPWM, as shown in figure 9, or to 2/3 of the cycle time for SVM.

(c) *Maximum inverter output in the controllable range:* Patterns that are generated on a three phase basis (SVM and SHE) maximize the output voltage for a given *dc* bus voltage; in SPWM, patterns are usually produced on a per leg basis, resulting in redundant switching on a line to line basis and a reduction in the output voltage. However, this value can be increased by injecting a third harmonic component in the modulating reference voltage.

6.3 *Output control*

The fundamental component of the output voltage of the inverter, $V_{i,1}$ (line to line peak), can be controlled, as required for var compensation purposes, figure 5c, either by changing the *dc* bus voltage V_{dc} or the PWM pattern, through modulation index control, M:

$$V_{i,1} = K_{ac} M V_{dc},$$

where K_{ac} is the gain of the converter, which depends upon the PWM pattern. For a three-phase inverter, the gain K_{ac} varies between 0.86 for SPWM, and 1.03 for SHE, SPWM with third harmonic injection and SVM.

The dynamic performance of the compensator depends largely on the control technique chosen.

(a) *DC bus voltage control:* If no PWM or a fixed SHE pattern designed for harmonic elimination only are used, output voltage control is obtained by varying the *dc* bus capacitor voltage V_{dc}. This is achieved by charging or discharging the *dc* capacitor; this can result in a slow response which depends upon the *dc* capacitor value, the *dc* voltage and the *ac* inductor (Joos *et al* 1991);

(b) *PWM pattern control:* If the PWM pattern is varied to produce a variable voltage from the fixed *dc* bus voltage, response can be very fast, in the order of the switching period. Instantaneous control of the injected current can then be implemented, and response is much faster than with conventional TCRs.

7. High power structures

In order to increase the power rating of the compensator to power system levels, typically 100 MVAR or more, available power switching devices must be combined in series and in parallel. However, advantage can be gained by combining converter units rather than individual devices: these units are associated so as to minimize harmonic injection while

maintaining switching losses low. This significantly increases compensator efficien
which must be kept very high (ideally 98% or higher).

Multi-stepped voltage waveforms that more closely approximate a sinusoidal volt
can be obtained in a number of ways, among which are multi-level and multi-pulse or mu
module structures. In addition to providing flexibility in the control of the fundamental a
harmonic components, these structures allow the increase of the voltage and current rat
to levels required for transmission systems. They also provide the option of remov
the coupling transformer used for voltage matching, which may reduce the cost of
installation (Hochgraf & Lasseter 1996). A number of structures are presented bel
based on voltage source inverters, and the implementation of harmonic cancellatio
discussed.

7.1   *Multi-pulse and multi-module structures*

Rather than increasing the frequency of the PWM pattern to reduce the harmonic conten
the output voltage, a number of units can be connected in series through a magnetic interf
circuit, figure 9b. Voltage distortion is reduced by harmonic cancellation or minimizati
Some multi-pulse structures use PWM. If no PWM is used, output voltage control mus
implemented by *dc* capacitor voltage control, as described in § 6.3 (Schauder *et al* 19
Agaki *et al* 1990).

In general, harmonic cancellation is achieved by phase shifting the harmonic compon
to be cancelled so that, when the output voltages of individual units are added, th
components cancel. This can be obtained by using the following.

(a) *Phase-shifted transformer windings:* Converter units are identically gated, figure
but the fundamental component of the output is shifted by $60°/n$ for n converters
applied to transformers with a phase shift of $60°/n$, figure 9b; this allows total harmo
cancellation (Mori *et al* 1993).
(b) *Patterns with phase-shifted harmonics:* Phase-shifted carrier techniques, as show
figure 8a for 8 units, exhibit this feature; these techniques however only provide harmo
minimization; for example, in figure 8b, individual converters have harmonic compone
around the 9th and multiples, whereas the combined output of 4 units, figure 8c, c
contains dominant components around the 36th and multiples (Kuang & Ooi 1994); tra
former·configurations are similar to that of figure 9b, but no phase shift is required
standard transformers can be used.

7.2   *Multi-level structures*

By splitting the *dc* bus capacitor, figure 10a, it is possible to generate a stepped vol
similar to that in figure 8b. A number of topologies are available among which the Di
Clamped Capacitor Multilevel Inverter (DCMLI), figure 10b and the Flying Capac
Multilevel Inverter (FCMLI) (Hochgraf *et al* 1994). Typically 3 to 7 level converters I
been investigated. The number of levels can be chosen to satisfy the waveform dister
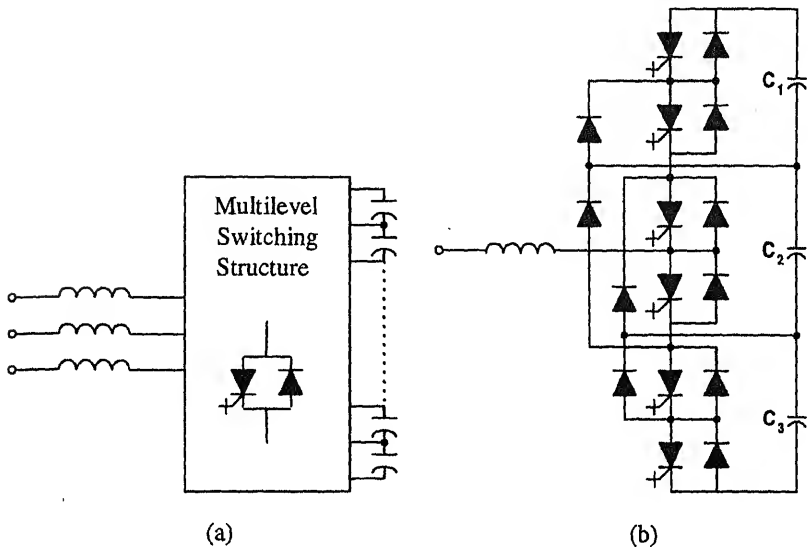
Figure 10. Multi-level compensator–voltage source topology. (**a**) General three-phase structure. (**b**) Power circuit for one-leg of a 4-level diode clamped converter.

7.3 *Single phase structures*

The structures presented above are based on three-phase converter units. This is the natural choice for three phase systems. However, implementations based on single-phase converters of the voltage source type have been proposed, figure 11 (Lai & Peng 1996; Peng & Lai 1996). Advantages include: (i) modularity and ease of expansion, obtained by stacking units, particularly to meet voltage ratings; (ii) control flexibility, since control is implemented on a per phase basis and unbalances are more easily handled. Output voltage waveforms are similar to that of figure 9b for a single pulse operation (Lai & Peng 1996). PWM can also be used.

8. Summary

Force-commutated static power compensators, part of the FACTS devices family and based on GTOs and on IGBTs, offer an interesting alternative to conventional shunt or series, fixed or variable, reactance compensation. They are capable of producing the required amount of reactive power independently of the line voltage or current: this is the result of their voltage or current source nature. They can both supply and absorb reactive power. Furthermore, as required, they can supply real power if equipped with energy storage devices. They have almost instantaneous response to control variable changes. Unlike systems based on passive components, they do not introduce potentially damaging resonant frequencies. Advanced static compensators are therefore one of the key elements in enhancing the
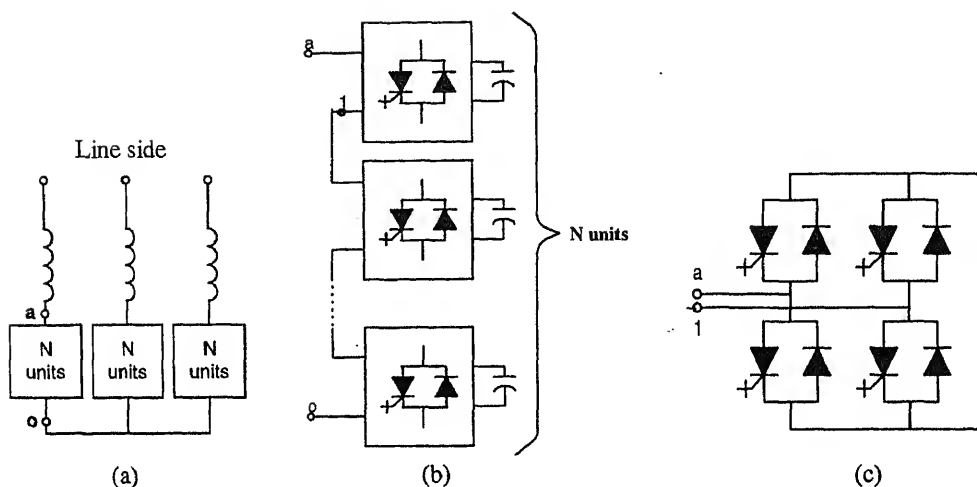
Figure 11. Multi-module compensator based on single phase units – voltage source topology. (a) Three-phase structure. (b) Multi-module single phase converter, series connection – one leg. (c) Circuit topology of each unit.

References

Agaki H, Kanazawa Y, Nabae A 1984 Instantaneous reactive power compensators comprising switching devices without energy storage. *IEEE Trans. Ind. Appl.* 20: 625–630

Agaki H, Tsukamoto Y, Nabae A 1990 Analysis and design of an active power filter using quad series voltage source PWM converters. *IEEE Trans. Ind. Appl.* 26: 93–98

Bakhshai A, Joos G, Jin H 1997 Space vector pattern generators for multi-module low switching frequency high power VAR compensators. In *IEEE Power Electon. Spec. Conf. PESC Rec.* vol. 1, pp 344–350

Campos A, Joos G, Ziogas P D, Lindsay J 1994 Analysis and design of a series voltage unbalance compensator based on a three-phase VSI operating with unbalanced switching functions. *IEEE Trans. Power Electron.* 9: 269–274

Edwards C W, Mattern K E, Stacey E J, Nannery P R, Gubernik J 1988 Advanced static VAR generator employing GTO thyristors. *IEEE Trans. Power Delivery* 3: 1622–1627

Espinoza J R, Joos G 1994 Three-phase series VAR compensation based on a voltage controlled current source inverter with supplemental modulation index control. In *IEEE Power Electr. Spec. PESC'94 Conf. Rec.* 2: 1437–1442

Gyugyi L 1979 Reactive power generation and control by thyristor circuits. *IEEE Trans. Ind. Appl.* IA-15: 521–532

Gyugyi L 1992 A unified power flow controller: A new approach to power transmission control

graf C, Lasseter R H 1996 A transformer-less static synchronous compensator employing ılti-level inverter. *IEEE Trans. Power Delivery* 11: 881–887

graf C, Lasseter R H, Divan D, Lipo T A 1994 Comparison of multilevel inverters for static r compensation. In *Conf. Rec. IEEE Ind. Appl. Soc. Annual Meeting*, pp 921–928

ᵢ, Joos G, Lopes L A C 1994 An efficient switched-reactor/capacitor-based VAR compen-tor. *IEEE Trans. Ind. Appl.* 30: 998–1005

G, Moran L, Ziogas P D 1991 Performance analysis of a PWM inverter VAR compensator. :*EE Trans. Power Electron.* 6: 380–391

›ri H A, Dewan S B, Lavers J D 1990 A large scale PWM solid state synchronous condenser. ᵢ *Conf. Rec. IEEE Ind. Appl. Soc. Annual Meeting*, pp 1099–1106

ng J, Ooi B T 1994 Series connected voltage-source converter modules for force-commutated VC and DC transmission. *IEEE Trans. Power Delivery* 9: 977–983

J-S, Peng F Z 1996 Multilevel converters – A new breed of power converters. *IEEE Trans. nd. Appl.* 32: 509–517

ˌsen E, Miller N, Nilsson S, Lindgren S 1992 Benefits of GTO-based compensation systems ᵒr electric utility applications. *IEEE Trans. Power Delivery* 7: 2056–2062

ꜱes L A C, Joos G, Ooi B T 1996 A multi-module switched-reactor-based static VAR compen-ꜱator. In *IEEE Power Electron. Spec. PESC'96 Conf. Rec.* 1: 515–520

nzies R W, Zhuang Y 1995 Advanced static compensation using multilevel GTO thyristor inverter. *IEEE Trans. Power Delivery* 10: 732–738

ɔran L, Ziogas P, Joos G 1989a Analysis and design of a 3-phase current source solid-state VAR compensator. *IEEE Trans. Ind. Appl.* IA-25: 356–365

ɔran L, Ziogas P, Joos G 1989b Analysis and design of a three-phase synchronous solid-state VAR compensator. *IEEE Trans. Ind. Appl.* IA-25: 598–608

ɔri S, Matsuno K, Hasegawa T, Ohnishi S, Tadeka M, Seto M, Murakami S, Ishiguro F 1993 Development of a large static VAR generator using self-commutated inverters for improving power system stability. *IEEE Trans. Power Syst.* 8: 371–377

ɔi B T, Dai S-Z 1993 Series-type solid state VAR compensator. *IEEE Trans. Power Electron.* 8: 164–169

ₑng F, Lai J.-S. 1996 Dynamic performance and control of a static VAR generator using cascade multilevel inverters. In *Conf. Rec. IEEE Ind. Appl. Soc. Annual Meeting*, pp 1009–1015

chauder C, Gernhardt M, Stacey E, Lemark T, Gyugyi L, Cease T W, Edris A 1995 Development of a ±100 MVAR static condenser for voltage control of transmission systems. *IEEE Trans. Power Delivery* 10: 1486–1493

ᵢuzuki H, Nakajima T, Izumi K, Sugimoto S, Mino Y, Abe H 1997 Development and testing of prototype models for a high performance 300 MW self-commutated AC/DC converter. *IEEE Trans. Power Delivery* 12: 1589–1601

√enkataramanan G, Johnson B 1997 A pulse width modulated power line conditioner for sensitive load centers. *IEEE Trans. Power Delivery* 12: 844–849

Walker L 1986 Force-commutated reactive-power compensator. *IEEE Trans. Ind. Appl.* IA-22: 1091–1104

Walker L 1990 10-MW GTO converter for battery peaking service. *IEEE Trans. Ind. Appl.* 26: 63–72

Zhang Z-C, Ooi B T 1993 Multi-modular current source SPWM converters for superconducting ___ ___ ___ ___ ___ *IEEE Trans. Power Electron.* 8: 250–256

tive power filters – Recent advances

NED MOHAN and GIRISH R KAMATH

Dept. of Electrical & Computer Engineering, 200, Union St. S.E., University
of Minnesota, Minneapolis, MN 55455, USA
e-mail: [mohan,gkamath]@ece.umn.edu

Abstract. Power electronic loads inject harmonic currents into the utility
causing overheating of power transformers and neutral wires, the power sys-
tem, unpredictable performance of protection systems etc. In addition, electric
resonances in such loads can also cause other undesirable phenomena like
voltage fluctuations, radio frequency interference (RFI) etc. To mitigate these
undesirable effects, a new class of power electronics equipment (Active Filters)
is being considered. A review of present-day solutions in the area of active fil-
ters is conducted in this paper. Finally, this paper discusses the trends in the
design of active filters and the factors influencing them.

Keywords. Active filters; harmonics; multi-level inverter; series-connection;
hybrid filters.

Introduction

le past few decades, there has been an increase in the use of electric power throughout
world. This increase has been accompanied by the use of equipment in our day to
life which has shown an increasing amount of sophistication and an appetite for good
ity electric power. The past few years has also witnessed the occurrence of a number
nishaps like distribution transformers catching fire, blackouts etc. Such incidents are
cting an increasing number of consumers as the use of electric power becomes more
espread. On investigating these accidents, it was found that one of the causes of these
urrences was electric power "pollution". This is generated by the use of modern ap-
nces and equipment connected to the grid. Typical examples at the residential level
ude computers, colour televisions, fluorescent ballasts etc. At the industrial level, high
er Adjustable Speed Drives (ASDs) and arc furnace loads are typical examples of
er polluting loads. These loads are used extensively in steel rolling mills, chemical
:ess industries etc.
:ommon undesirable characteristics of all these loads are the following.

They draw non-sinusoidal current from the utility, and

distort the utility voltage waveforms.
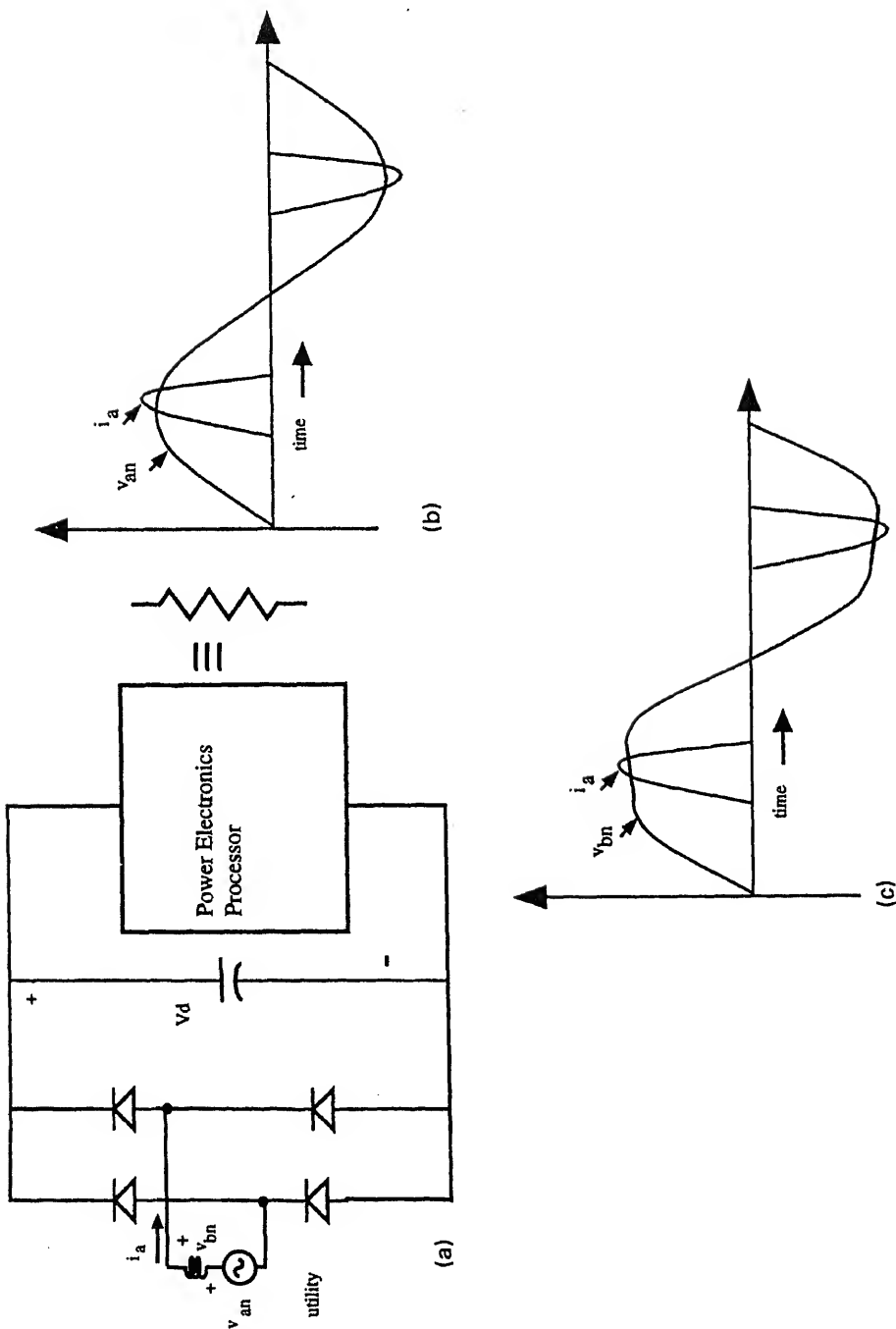
Figure 1. Equivalent circuit of a nonlinear load connected the utility. (a) Equivalent circuit of a nonlinear load. (b) Input

Table 1. Details of harmonic currents drawn by the load of figure 1.

Harmonic number	I_h/I_1 (in per unit)
3	0.77
5	0.42
7	0.15
9	0.08

From the point of view of studying the effect that such loads have on the utility, it possible to represent any of the above mentioned examples in terms of the schematic agram shown in figure 1a which consists of a diode bridge rectifier and a resistor.

In this diagram, the power electronics processor is represented as an equivalent resistor. he current drawn by the load from the utility is shown in figure 1b. The voltage v_{bn} at the pint of common coupling (PCC), shown in figure 1c, is distorted due to the nature of the urrent drawn by the load. Thus any other consumer who is connected across points $b - n$ ould have to face this distorted voltage. The current drawn by the load has a dominant third armonic component as shown in table 1 which contains a table containing details of the urrent harmonics. The usage of such loads can lead to overheating of power transformers nd neutral wires, electric resonances in the power system, unpredictable performance of rotection systems etc. In addition, such loads can also cause other undesirable phenomena ke voltage fluctuations, flicker, sags, radio frequency interference (RFI) etc. (Redl *et al* 997).

In order to address this issue, certain institutions like the IEEE in the US and IEC (fol-wed in Europe) have established guidelines which set limits on the amount of distortion urrent that the load can inject into the utility. However, there is a difference in philosophy 1 the framing of guidelines between the two institutions.

The revision of IEEE-519 guidelines suggested by Duffey & Stratford (1988) became fficial in 1992. It takes into account the relative strength of the utility system and sets mits on the amount of harmonics (as a percentage of the fundamental current) that a user an inject at the Point of Common Coupling (PCC). In the case of the IEC-555 standard hich was issued in the early 1980s to suggest guidelines for single-phase applications, he emphasis is on limiting the absolute amount of harmonics that is generated at the quipment level. The other major difference is that IEC standards have been enforced into w while IEEE 519-1992 guidelines are still recommendations.

Active filtering and current waveshaping techniques

1 order to address the issue of harmonic current injection, there are two options.

1) Incorporate current waveshaping circuits within the equipment so that they draw sinusoidal line currents.

2) Install filters to "clean up" the distorted current waveforms drawn by these loads.
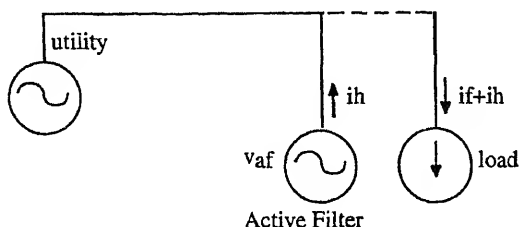
Figure 2. Single line diagram for active filter acting as a harmonic current source.

not be the most cost-effective option, especially at medium to high power level (great than 10 kW).

In such cases it may be more economical to follow the second approach by installi an active filter unit at the plant level in order to solve the problem.

(1) *Current waveshaping techniques:* IEC standards limit the current harmonics that a generated by single-phase electrical appliances. This encourages the approach whe one inserts a power electronics block between such a load and the utility to e sure that it (the utility) is insulated from the effects of the nonlinear load, i.e. t diode bridge rectifier of figure 1a is replaced by a "utility friendly" block whi draws sinusoidal current from the mains at unity power factor to maximize efficienc It then generates the required DC voltage at which the power electronic process operates.

(2) *Active filtering techniques:* IEEE 519-1992 recommendations limit the injection current harmonics by a user at the Point of Common Coupling (PCC). For examp the IEEE 519-1992 recommends a limit of 5% in weak systems on the Total Harmor Distortion (THD) in the current at the Point of Common Coupling. The total harmor distortion (THD) in the current i_s drawn from the utility is the ratio of the *rms* val of its distortion component to the *rms* value of its fundamental-frequency compone: It is given as,

$$\% \text{ THD} = \left[\left(\sum_{h=2}^{\infty} I_h^2 \right) \right]^{1/2} \bigg/ I_1 \times 100$$

where

$I_h = rms$ value of the current at harmonic order h,

$I_1 = rms$ value of the fundamental-frequency current component.

In this case, the user is free to use any equipment in his premises which are not necessar utility "friendly" as long as he uses external harmonic current reduction techniques meet the guidelines specified. Active filters fall in this category. They are harmonic curre sources which are used in conjunction with the utility to provide the required harmor
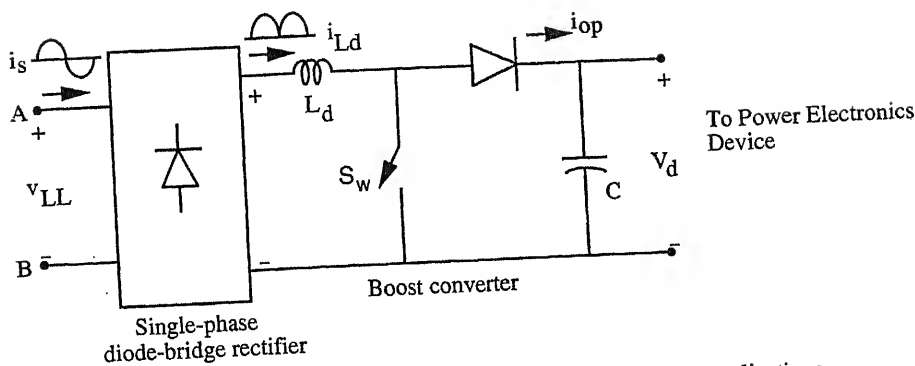
Figure 3. Current waveshaping technique for single-phase applications.
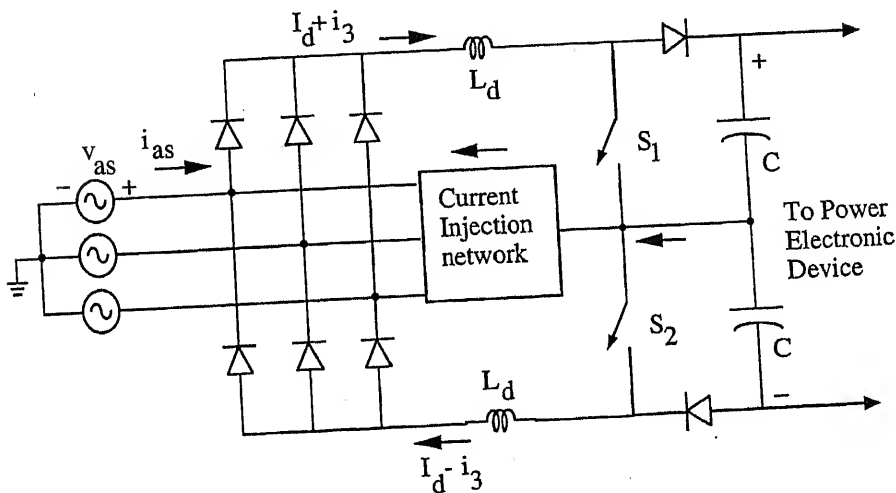
Figure 4. Current waveshaping technique for three-phase applications.

2.1 *Current waveshaping techniques*

In this section, some of the waveshaping techniques used for single and three-phase applications are reviewed. Figure 3 shows a typical current waveshaping technique for single phase loads.

    The circuit is a boost converter drawing unity power factor current from the utility. The output DC voltage V_d is regulated to a value which is higher than the peak of the utility voltage v_{LL}. In the boost converter approach, the input current i_s is in phase with v_{LL}. The power electronics processor is then connected to this output. This method is extremely popular for single-phase applications. For three-phase applications, a current waveshaping technique is shown in figure 4. This method is applicable where input transformer isolation is required for other reasons. In this method, a third harmonic current is injected through the current circulation network reducing the THD of the current drawn from the utility. The principle of operation, design and implementation are discussed by Naik (1993) and Rastogi (1993). The output voltage is regulated to be at a level higher than the peak of the

input voltage. For those cases where input isolation is not required, a topology called
Vienna Rectifier may be used (Kolar & Zach 1994). The method described by Kolar *et*
(1997) can be used for universal input three-phase voltages or where there are large in
voltage variations.

2.2 *Passive filters*

Filters are a method of "cleaning" up the current waveforms drawn from the utility in ca
where the load does not contain a waveshaping circuit. Active and passive filters are u
either together (to form hybrid filters) or separately (to reduce current harmonics).

Passive filters can be connected either in parallel with the load (shunt filter) or in ser
with it (series filter). An ideal shunt filter serves as a short circuit for load harmo
currents. A series filter consists of an impedance connected in series with the utility a
acts as an open circuit to load harmonic currents. The disadvantages with passive filt
are as follows.

(1) They are bulky;

(2) the filter impedance can resonate with the AC system impedance, resulting in increa
    voltage distortion and damage to equipment;

(3) overcompensation of Vars can result in lowering the power factor.

These drawbacks of passive filters can be overcome by active filters.

2.3 *Active filters*

Active filters can also be connected in series and shunt mode. An example of an act
filter for a three-phase, four-wire system (Quinn *et al* 1993) is shown in figure 5.

Equivalent circuit of a hybrid filter
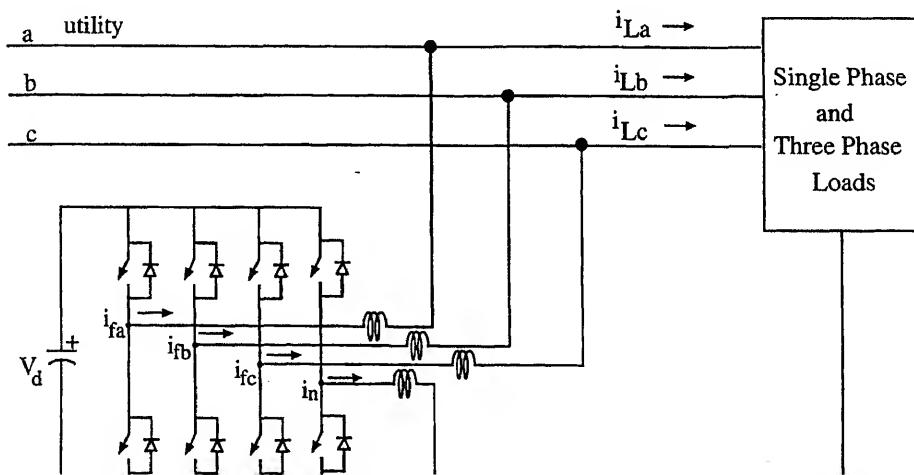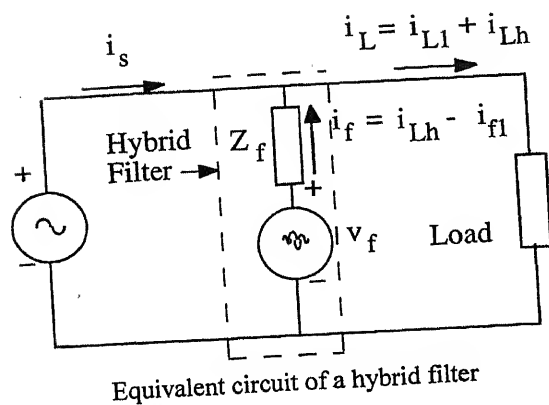
Figure 6. Equivalent circuit of a hybrid filter.

Here, the four-legged converter (optionally through a line frequency transformer) is connected to the utility. The switch-mode converter neutralizes the load harmonic currents in all phases, including the neutral), such that the utility supplies only the line-frequency currents. This topology is well suited at low and medium power levels where the required space needs to be minimized. The VA rating of the converter switches is very high. This scheme is not suitable at higher power levels (greater than 50 kW) as operation of the converter at lower switching frequency leads to a lower current loop band-width and hence it is not possible to neutralize higher frequency current harmonics. Another concern is the EMI (electro-magnetic interference) produced by such an arrangement.

2.4 Hybrid filters

Hybrid filters combine the ruggedness of passive filters with the above mentioned advantages of active filters. Figure 6 is an example of a hybrid filter (Rastogi *et al* 1995). In this scheme, the active portion of the filter (shown as a voltage source v_f) is connected in series with a passive impedance Z_f. Z_f is used to block the utility voltage by making v_f draw a fundamental frequency current i_{f1} through it. This reduces the voltage rating and hence the kVA rating of v_f substantially. However, the design of this filter is dependent on system parameters and knowledge of the load harmonic currents are required in order to design Z_f. Also, hybrid filters are not as compact as all active filters.

2.5 All solid state per-phase solution for utility applications

Often in utility-interactive applications at high power, the utility voltage is of the order of a few kilovolts, for example, 13.8 kV or 34.5 kV. To meet the voltage requirement with devices of limited voltage ratings (3 to 4 kV as with GTOs), the following options are available.
(1) A line frequency voltage matching transformer to translate the utility voltage level to a value at which the devices can operate. This is the simplest solution used almost always in the past. However, there are several problems associated with this approach. The line-frequency transformer is bulky and causes significant losses and, being a
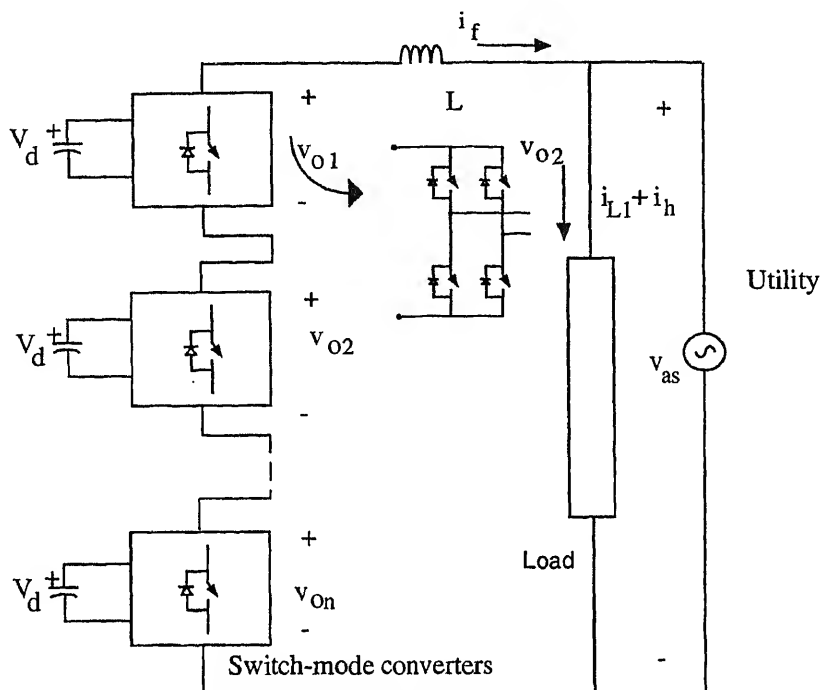
Figure 7. Schematic diagram of a hybrid all-switching series converter for utility applications.

mature technology, its cost is not expected to decline. The presence of a line-frequency transformer also results in the need for an additional DC demagnetizing current control loop to prevent the transformer from becoming saturated due to non-idealities in the main control loop and in the power circuit.

(2) Multi-level inverters are being considered (Baker 1980; Hochgraf *et al* 1994; Matsui 1995). However, for levels greater than three, the current flow in and out of the neutral point results in voltage variations of the neutral. This calls for additional neutral point voltage control (Fukuda & Sagawa 1995; Matsui 1995) and a limitation in the switching frequency.

(3) The switches in the switch-mode converter may comprise of a series connection of discrete devices (e.g. GTOs). However the devices are derated and operate below their rated voltage rating (Takeda *et al* 1995; Ichikawa *et al* 1995).

A solution to the above problems, as shown in figure 7, is proposed by Paice & Edwards (1987), Mohan & Kamath (1995) and Peng *et al* (1995). It consists of a number of converters whose outputs are connected in series, on a per-phase basis. The arrangement has the following properties: It consists of a set of slow switching converters (switching at line frequency or its low multiple) whose outputs are connected in series with a fast-switching PWM converter. The DC bus voltages in these converters are independent of each other and may or may not be equal. The devices are matched to their switching requirements, for

block the utility voltage, while that of the fast switching converter is to inject the
priate current. Since the fast switching converter operates at low voltage and hence
power, it produces much lower EMI (electro-magnetic interference) as compared to
le conventional converter operating at high power. There is also a decrease in the
f the interface inductor as compared to a single six-switch converter topology for the
current ripple and switching frequency.

ulation and experimental results of a proof of concept prototype with three converters
cted in series and tested for an active filter application are presented by Mohan &
th (1997). With improvements in device technology, both in terms of performance
st, and the use of digital control techniques to improve dynamic performance of the
, there is a trend towards using all-solid state solutions to power quality problems.
s consistent with the PEBB (power electronics building block) concept where the
switches are combined with their gate drives and associated control circuitry into
le module, for a variety of applications. There have been continuing advances in
onductor technology resulting in improvements in device switching characteristics
er voltages, especially with the emergence of SiC (silicon carbide) devices (Weitzel
996). This also indicates that the trend will be towards solid-state solutions for power
onic applications in general and active filters in particular.

desire for better quality power also means an increase in the complexity of the
ons that will have to be performed by the active filters of the future. Akagi (1996)
oposed the concept of a general unified power quality conditioner which consists
eries active filter and a shunt active filter. This power conditioner, in addition to
nic compensation, also takes care of voltage regulation, voltage flicker/imbalance
ensation, negative sequence current compensation and electric resonance damping.

onclusions

has been an increase in the use of power electronic loads causing concern for power
y. This paper describes some of the methods that are currently being considered to
ss this issue.

ences

H 1996 New trends in active filters for power conditioning. *IEEE Trans. Ind. Appl.* 32:
2–1322
R 1980 High-voltage converter circuit. *US Patent No. 4,203,151*
y C, Stratford R 1988 Update of harmonic standard IEEE 519: IEEE recommended practices
l requirements for harmonic control in electric power systems. *IEEE Ind. Appl. Soc. (PCIC)*
nference

Ichikawa F, Yajima M, Nakajima T, Irokowa S, Kawakami N 1995 Operating experience of a 50 MVA self-commutated SVC at the Shin-Shinano substation. *IPEC-Yokohama 1995 Conference Records.* vol. 1, pp 597–602

Kolar J, Zach F 1994 A novel three-phase utility interface minimizing line current harmonics of high-power telecommunications rectifier modules. *IEEE ITEC Conference Proceedings*, pp 367–374

Kolar J, Hari S, Drofenik V, Mohan N, Zach F 1997 A novel three-phase three-switch three-level high power factor SEPIC-type AC-to-DC Converter. *IEEE APEC Conference Proceedings*, pp 657–665

Matsui M 1995 Method for controlling neutral point for static var compensation. *IPEC-Yokohama 1995 Conference Records*, vol. 1, pp 488-493

Mohan N, Kamath G R 1995 A novel, per-phase interface of power electronic apparatus for power system applications. *North American Power Symposium 1995 Proceedings*, pp 457–461

Mohan N, Kamath G R 1997 A hybrid all-switching per-phase solution for power electronics utility application. *IEEE IECON Conference Proceedings* (to be published)

Naik R 1993 *A novel sinusoidal-current, 3-phase utility interface: hardware implementation.* M S thesis, University of Minnesota, Minneapolis

Paice D, Edwards C 1987 High voltage modular inverter and control system thereof. *US patent No. 4674024*

Peng F, Lai J, McKeenver J, Van Coevering J 1995 A multilevel voltage-source inverter with separate DC sources for static var generation. *IEEE-Ind. Appl. Soc. Conference Records*, pp 2541–2548

Quinn C, Mohan N, Mehta H 1993 A four-wire, current-controlled converter provides harmonic neutralization in three-phase, four wire systems. *IEEE-APEC Conference Proceedings*, pp 841–846

Rastogi M 1993 *Analysis and optimization of a novel 3-phase, sinusoidal line current rectifier.* M S thesis, University of Minnesota, Minneapolis

Rastogi M, Mohan N, Edris A 1995 Hybrid-active filtering of power system harmonics. *IEEE Trans. Power Delivery* 10: 1994–2000

Redl R, Tenti P, van Wyk J 1997 Combatting the pollution of the power distribution systems by electronic equipment. *IEEE Appl. Power Electron. Conf. Exposition*, pp 42–48

Takeda M, Murukami S, Iizuka A, Hirakawa M, Kishida M, Hase S, Mochinaga H 1995 Development of an SVG series for voltage control over three-phase unbalance caused by railway load. *IPEC-Yokohama 1995 Conference Records.* vol. 1, pp 603–608

Weitzel C, Palmour J, Carter C, Moore K, Nordquist K, Allen S, Thero C, Bhatnagar M 1996 Silicon carbide high-power devices. *IEEE Trans. Electron. Devices* 43: 1732–1739

gh power factor operation of resonant converters on the lity line

A K S BHAT and V BELAGULI[*]

Department of Electrical and Computer Engineering, University of Victoria, Victoria (BC), V8W 3P6, Canada
Present address: Department of Electrical Engineering, Singapore Polytechnic, Singapore 139 651
e-mail: bhat@ece.UVic.CA; belaguli@sp.ac.sg

Abstract. Operation and characteristics of resonant converters on the utility line are presented. Series-parallel (LCC-type) resonant converter operating with discontinuous current mode and continuous current mode (variable frequency control as well as fixed-frequency) are considered. Design examples are presented. SPICE simulation and experimental results obtained for the designed converters (rated at 150 W) are presented to verify the theory. It is shown that high line power factor (>0.95) and line current total harmonic distortion (THD) of $<25\%$ are obtained for the LCC-type converter for a wide load range (from full load to 10% rated load) without any active control, and the switch peak current decreases with the load current. With active line current control, low distortion and zero voltage switching for the entire cycle are realized.

Keywords. Resonant converters; power factor; zero-voltage-switching; zero-current-switching.

Introduction

ring the last few years, high-frequency (HF) soft-switching (including resonant) con-ters have gained popularity due to their advantages compared to hard-switching PWM verters. Some of the advantages are: zero-voltage-switching (ZVS) or zero-current-tching (ZCS) reducing switching losses, high efficiency, higher switching frequency, all size, low weight etc. The ZVS or ZCS reduces the switch stresses during the switch -on or turn-off instants. Many of the configurations utilize the leakage inductance of HF transformer as part of the resonant inductor eliminating the voltage spike problems sent in PWM converters.

With an increase in the number of power converters operating on the utility line, har-nics generated by these converters have become a matter of great concern. Due to the orcement (either already existing or being proposed) of strict harmonic regulations

(e.g. IEC555), there is an increased interest in the reduction or elimination of line current harmonics. The methods proposed or implemented in the literature to reduce the line current harmonics can be classified as:

(1) use of line filters (active or passive);

(2) use of power converters with active control of line current (Kataoka *et al* 1979; Kocher & Steigerwald 1983; He & Mohan 1987);

(3) modifications to the existing converters (Schlecht & Miwa 1987; Keraluwala *et al* 1991; Schutten *et al* 1991). The last method is the simplest solution.

Advantages of HF resonant converters can be utilized in *ac*-to-*dc* converters (Chambers 1983; Nijhof 1986; He & Mohan 1987; Schlecht & Miwa 1987; Keraluwala *et al* 1991; Schutten *et al* 1991; Belaguli & Bhat 1995; Belaguli 1996) for realizing power conversion with improved performance (high efficiency, high power factor and low line current harmonics etc.) while reducing the size, weight and cost. Perhaps the earliest work in this area was reported by Chambers (1983). The series resonant converter (SRC) operating in discontinuous current mode (DCM) was used without any line current control. Due to forced shut off of the SRC when the line voltage is around 50% of its peak value, the maximum power factor achievable and reduction in total harmonic distortion (THD) are limited. Active line current wave shaping has been used in a single ended resonant converter (He & Mohan 1987).

2. Resonant converters and their operation on the utility line

A number of resonant converter configurations are realizable using different resonant tank circuits (Steigerwald 1988; Bhat 1991) and the three most popular configurations are: series resonant (or series loaded) converter (SRC), parallel resonant (or parallel loaded) converter (PRC) and series-parallel (or LCC-type) resonant converter (SPRC). Load voltage regulation in such converters for input supply variations and load changes is achieved by either varying the switching frequency or using fixed-frequency (variable pulse-width) control. The SRC has high efficiency from full load to part load. But it requires wide variation in switching frequency for power control and the output filter capacitor must carry high ripple current when used in low output voltage and large load current application. The PRC is suitable for low output voltage with large load current applications and it requires narrow variation in switching frequency for power control. However, peak currents through the switches do not decrease with load current reducing the efficiency at reduced loads. The SPRC takes the desirable features of both SRC and the PRC.

Keraluwala *et al* (1991), Schutten *et al* (1991) and Belaguli & Bhat (1995) have reported the operation of resonant converters with high power factor. But the line current harmonic distortion was very high (18% minimum around half-load to about 48% at full-load) and the switch peak current did not decrease with the load current since PRC was used (Schutten *et al* 1991). Use of LCC-type converter had been discussed briefly. The major problem was the requirement of keeping the ratio of switching frequency to series resonance frequency

ventionally, HF link dc-to-dc resonant converters were operated with a smooth
ut. This means, when the converter has to be operated on the utility line, a line
r followed by a large filter capacitor (C_i) has to be used to get a dc supply. The
antages of this approach are well known: High peak currents drawn from the line with
harmonic distortion (THD) greater than 80% and low line pf (<0.65). However,
eal single-phase ac-to-dc converter, the input line current has to be sinusoidal and
e with the line voltage. This results in power with a double line frequency ($2f_L$,
f_L is the line frequency) having a peak value of $2P_o$, where P_o is the average output

e proposed ac-to-dc converters, a small HF filter capacitor C_i is used at the input
en et al 1991; Belaguli & Bhat 1995; Belaguli 1996). The dc link voltage in this
approximately a rectified sine wave. This allows the HF inverter to draw current
e utility line over the entire cycle reducing the peak currents. However, a large
ive output filter C_d is required to filter the 120 Hz voltage ripple at the output to
a constant voltage V_o. Due to the rectified sine-wave at the input of the HF inverter,
rter output voltage v_{ab} is varying and the instantaneous normalized converter gain
$= V'_o/v_s(t)$ must also vary in order to draw sinusoidal line current and maintain
ant dc output voltage. The voltage gain $M(t)$ is minimum at the peak of the ac
e, and maximum near the zero crossings of the ac voltage.
closely the converter gain matches the required gain variation in order to get a line
which follows the input sine wave depends on the resonant converter configuration.
ase of SRC, the converter does not have enough gain for most part of the line cycle,
sulting in poor pf and large line current distortion. It is well known that PRC has
ain near valley points of line voltage at full load (for switching frequencies near
nce) due to its high Q (for the PRC, $Q = R_L/(L_r/C_p)^{1/2}$, where L_r and C_p are the
nt inductance and capacitance, respectively) near zero crossings. This is because,
varies along the ac cycle, with maximum loading at the peak of ac and light loading
ero crossings of ac cycle. In fact, the gain obtained is much higher than required
g the line current to approach a square-wave resulting in a higher THD of line current
load. In addition, the switch peak currents are high in the case of PRC and does not
se with the load current. Only limited discussion on the use of LCC-type converter
ven (Schutten et al 1991) due to insufficient gain available at the valley points of
ltage, and the ratio of switching frequency to series resonance frequency was kept
$\simeq 1.51$) to maintain lagging pf (above resonance) mode of operation resulting in
switch peak currents. But our research has shown that LCC-type converters can
rated with high pf and low harmonic distortion while maintaining its advantages.
ore, an LCC-type converter is taken as the example in this paper.

peration of SPRC on the utility line

Emulation of a resistor

1 shows the SPRC (or LCC-type resonant converter) operating on the utility line
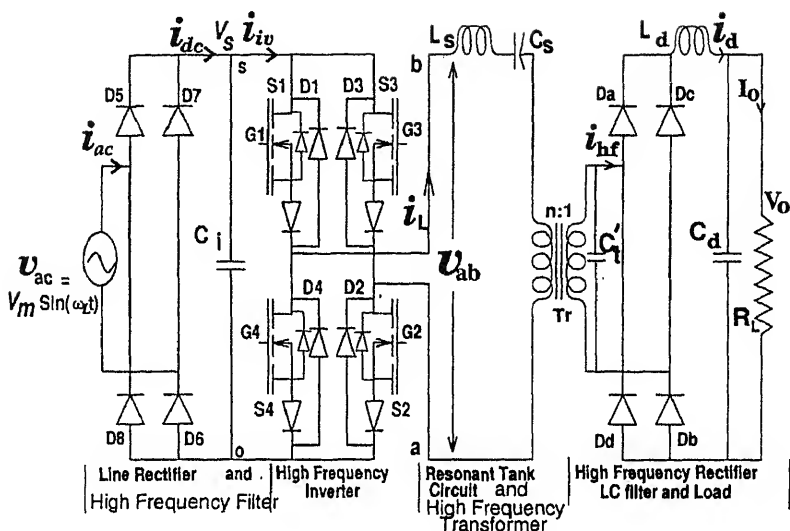
Figure 1. High-frequency transformer isolated *ac*-to-*dc* converter employing series-parallel resonant converter. Note that C_i is an HF filter capacitor.

is a rectified sine-wave. The output *dc* link filter inductor current is given by

$$i'_d(t) = I'_{dm} \sin^2(w_L t)$$

where

$$I'_{dm} = 2P_o/V'_o, \quad V'_o = nV_o.$$

Also, series resonant $Q_s(t)$ expressed in terms of time varying reflected load resistan
achieves its maximum at the peak of the line voltage cycle, and is given by

$$Q_s(t) = w_s L/R'_L(t) = Q_{s\,\max} \sin^2(w_L t)$$

where

$$Q_{s\,\max} = (L/C_s)^{1/2}/R'_{Lp} = (L/C_s)^{1/2}(I'_{dm}/V'_o),$$

$R'_{Lp} = V'^2_o/(2P_o)$, $M = V'_o/V_m$, $L = L_s + L_l$, L_s is the external resonant inductan
L_l is the leakage inductance of the HF transformer.

Also, define: $y_s = 2\pi f_s/\omega_{sr}$, $\omega_{sr} = 1/(LC_s)^{1/2}$, f_s is the switching frequency.

To get a sinusoidal line current, that is, for resistive emulation, the converter gain
to be changed by exercising active control. However, it is shown in later sections that,
matching the converter operating characteristics using the design constraints (§ 3.3),
SPRC can be operated on the utility line, to obtain high *pf* with low harmonic curr
distortion even without active control.

3.2 *Operating modes of a SPRC*

When variable frequency operation is used, depending on the switching frequency, SP
can operate in discontinuous current mode (DCM) or continuous current mode (CC

Design constraints

following are the design criteria to operate the resonant converters on the utility line high *pf* and low line current harmonic distortion.

$Q_{s\,max}$ should be chosen to minimize the inductor rms current and also the kVA/kW ating of the resonant tank circuit. Higher $Q_{s\,max}$ minimizes inductor rms current, but ncreases its size.

Depending on the type of control used for regulating the *dc* output voltage from full oad to reduced load, the range of variation in (a) switching frequency required in ase of variable frequency DCM or CCM operation, or (b) phase shift (pulse width) equired in case of fixed-frequency CCM operation, must be minimized while choosing he C_s/C_t ratio to account for variation in input voltage in addition to the required gain *M*.

The switching frequency (or pulse width) should be chosen such that the resonant onverter operates only in the modes for which it is designed for full load near the peak of the supply voltage.

The resonant converter generates the required gain at least from 30° to 150° (as most of the total output power is delivered in this range) even without active control over he 60 Hz (or 50 Hz) *ac* half cycle, to get lower line current distortion.

Discontinuous current mode operation

converter is designed to operate in just continuous current mode (JCCM) (figure 2a) the rated minimum peak of *ac* line voltage with rated maximum load current and peration will be in DCM (figure 2b) for the rest of line cycle. The output voltage is lated by decreasing the switching frequency such that the converter operates in DCM. figure 2a when S1 and S2 are turned on, resonant current j_L (normalized i_L) starts ng with positive polarity of the voltage v_{ab} applied to the tank circuit. During interval-m_{ct} (normalized v_{ct}) $= 0$ and the load is supplied by the resonant current until $= J(t)$ (normalized load current assumed to be constant over one switching half) to enter interval-A. In interval-A, the resonant current in excess of instantaneous load nt $J(t)$ charges the parallel capacitor C_t. The gating pulse to the switch is removed, the diodes D1 and D2 start conducting the resonant current in interval-A due to sal in current polarity. The diodes D1 and D2 turn off with zero current at the end terval-C2. During the additional interval-E (dead gap) shown in figure 2b, the input ly to the tank circuit is cut off as all the switches are in the OFF state. It must be noted the series capacitor voltage remains constant during interval-E (same as the voltage e end of interval-C2), until the other pair of switches S3 and S4 are turned ON. The tion of interval-E increases with decrease in load current. The operation of converter is lar for the other half cycle, where switches S3 and S4 are turned ON, except for change larity in voltages and currents. Note that, all the high frequency rectifier diodes in the ut section are in conduction during intervals-C1 and C2.

the output section, the filter components L_d and C_d carry approximately 120 Hz cur- and voltage ripple respectively, under high power factor operation. The filter inductor
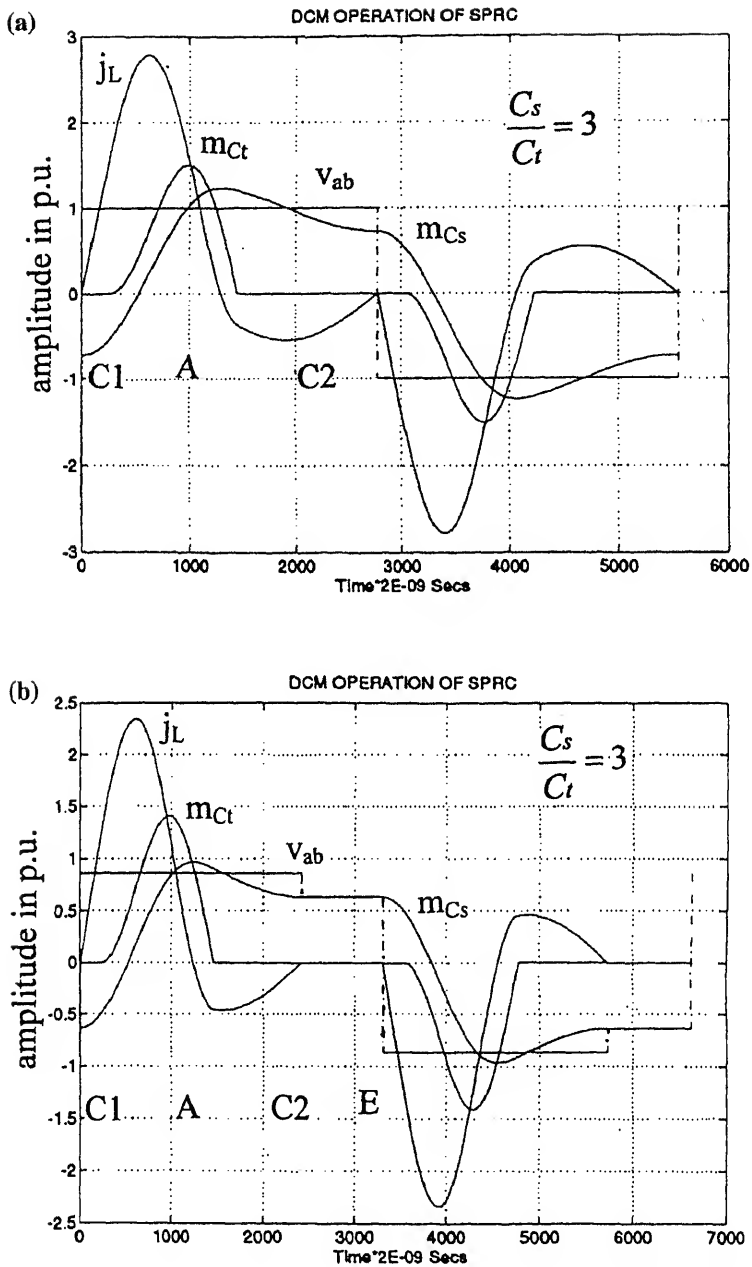
Figure 2. Normalized steady-state operating waveforms obtained from PROMATLAB for SPRC operation in DCM. **(a)** Just CCM operation. **(b)** DCM operation.
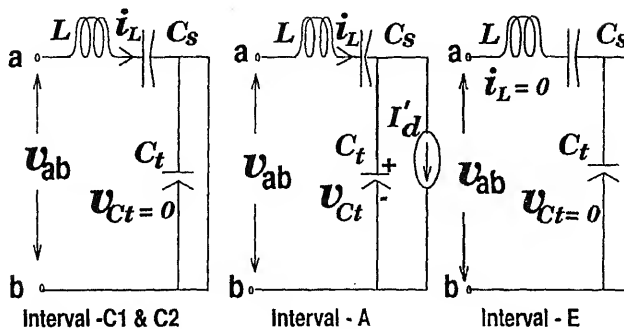
Figure 3. Equivalent circuit models for SPRC operating in DCM.

to filter only HF ripple ($2 f_s$, where f_s is the switching frequency), whereas, C_d
ilter the 120 Hz ripple.

models required to analyse the converter during different intervals of operation
ied as interval-C1, A, C2, and E) for the waveforms of figure 2 are shown in
. In view of low switching current ripple in $i_d(t)$ of the output section, constant
model has been used. Based on the above equivalent circuit models, the converter
en analysed using the state-space approach (it is complex due to four intervals
ation and three state variables) (Belaguli 1996). If active control is used to get
dal line current, the variation in switching frequency over the 60 Hz *ac* half cycle
given instantaneous $M(t)$ and $Q_s(t)$, have been evaluated using the converter gain
n. Figure 4 shows the plot of $Q_s(t)$, $M(t)$, and $y_s(t)$ over the 60 Hz *ac* half cycle,
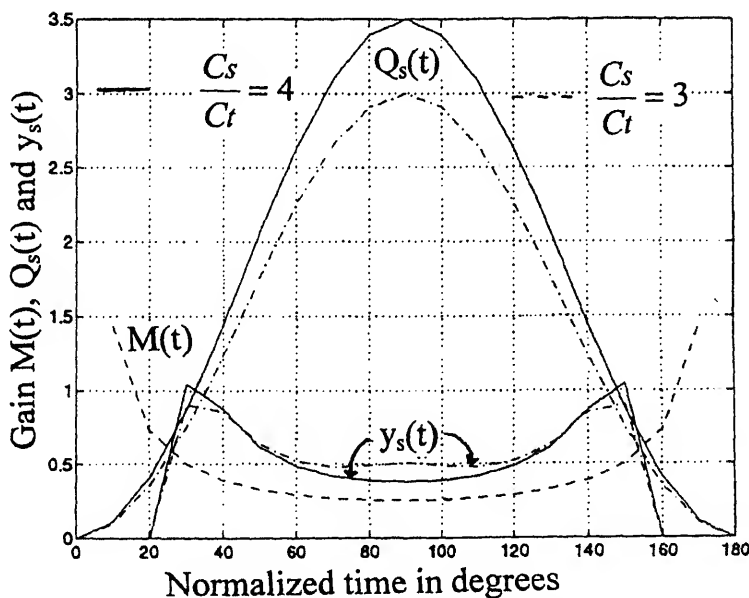
igure 4. Variation of series $Q_s(t)$, required converter gain $M(t)$, and normalized switch-
g frequency ratio y_s to get sinusoidal line current over *ac* half cycle for minimum *ac*
ltage and rated maximum load conditions.

for two different values of $Q_{s\,max}$ and C_s/C_t ratio. However, if no active control is used, then it is necessary to properly choose the converter gain $M = V_o'/V_m$ and $Q_{s\,max}$ at the peak of the 60 Hz sinusoidal ac voltage cycle for a given C_s/C_t ratio.

From the analysis, it was found that, for a given peak converter gain of $V_o'/V_m = 0.25$, good compromised design values are $Q_{s\,max} = 3$ and $Q_{s\,max} = 3.5$ for C_s/C_t ratios of 3 and 4 respectively, with the converter operating in JCCM. But simulations showed that a capacitor ratio of 4 is preferable over 3, due to lower harmonic distortion and lower inverter output peak currents.

3.4a *Design example:* A converter is to be designed with the following specifications.
  Average output power, $P_o = 150$ W;
  Input voltage, $V_{ac} = 120$ to 135 V rms;
  Output voltage, $V_o = 48$ V;
  Full load current, $I_o = 3.125$ A;
  Output current ripple in i_d, $A_i = \pm10\%$ of I_o;
  Output voltage ripple, $A_v = \pm0.5\%$ of V_o;
  Series resonant frequency, $f_{sr} = 135$ kHz.
  The design calculations are done for SPRC delivering a peak power of $2P_o$ by choosing the Q_s at the peak of the ac line cycle (JCCM operation at this point) for rated minimum input voltage, i.e., $V_{ac} = 120$ V.
  Using $M = 0.25$, $V_m = \sqrt{2} \times 120$, $V'o = MV_m = 42.5$ V.
  $R_{Lp}' = V_o'^2/(2P_o) = 6.02\,\Omega$, $y_s = f_s/f_{sr} = 0.4$, transformer turns ratio $n:1 = 0.8854$.
  Using the definitions of $Q_{s\,max}$ and y_s, the following component values are obtained for $Q_{s\,max} = 3.5$ and $C_s/C_t = 4$:

$$L = 24.84\,\mu\text{H}, \quad C_s = 0.0559\,\mu\text{F}, \quad C_t = 0.0139\,\mu\text{F}.$$

The output filter L_d and C_d are designed using the relationship given by Schutten *et al* (1991) and Belaguli (1996) to meet the output ripple requirements.

$$L_d = V_o/(3\pi f_s I_o A_i) = 150\,\mu\text{H}, \quad C_d = I_o/(180\pi V_o A_v) = 11,512\,\mu\text{F}.$$

3.4b *SPICE simulation results:* The 150 W, 48 V output converter designed above has been simulated in SPICE to verify its performance. Due to storage limitations, SPICE simulation studies have been done for a converter redesigned with a switching frequency of 25 kHz, which does not change the operating principle and the actual results. Figure 5 shows the various waveforms obtained from SPICE simulation for the DCM operation of SPRC, for different load currents and $C_s/C_t = 4$. The line current waveform for full load shown in figure 5a has a harmonic distortion of 14%. The JCCM operation of the converter near the peak of ac voltage waveform is shown in figure 5b. For 50% and 10% of the rated load, the line current waveforms are shown in figures 5c and 5d, which have distortion figures of 16% and 11% respectively, while the converter is operating only in the DCM mode. SPICE simulation studies showed that the line current THD for a capacitance ratio 3 was higher compared to 4. The switching frequency harmonic component appearing in
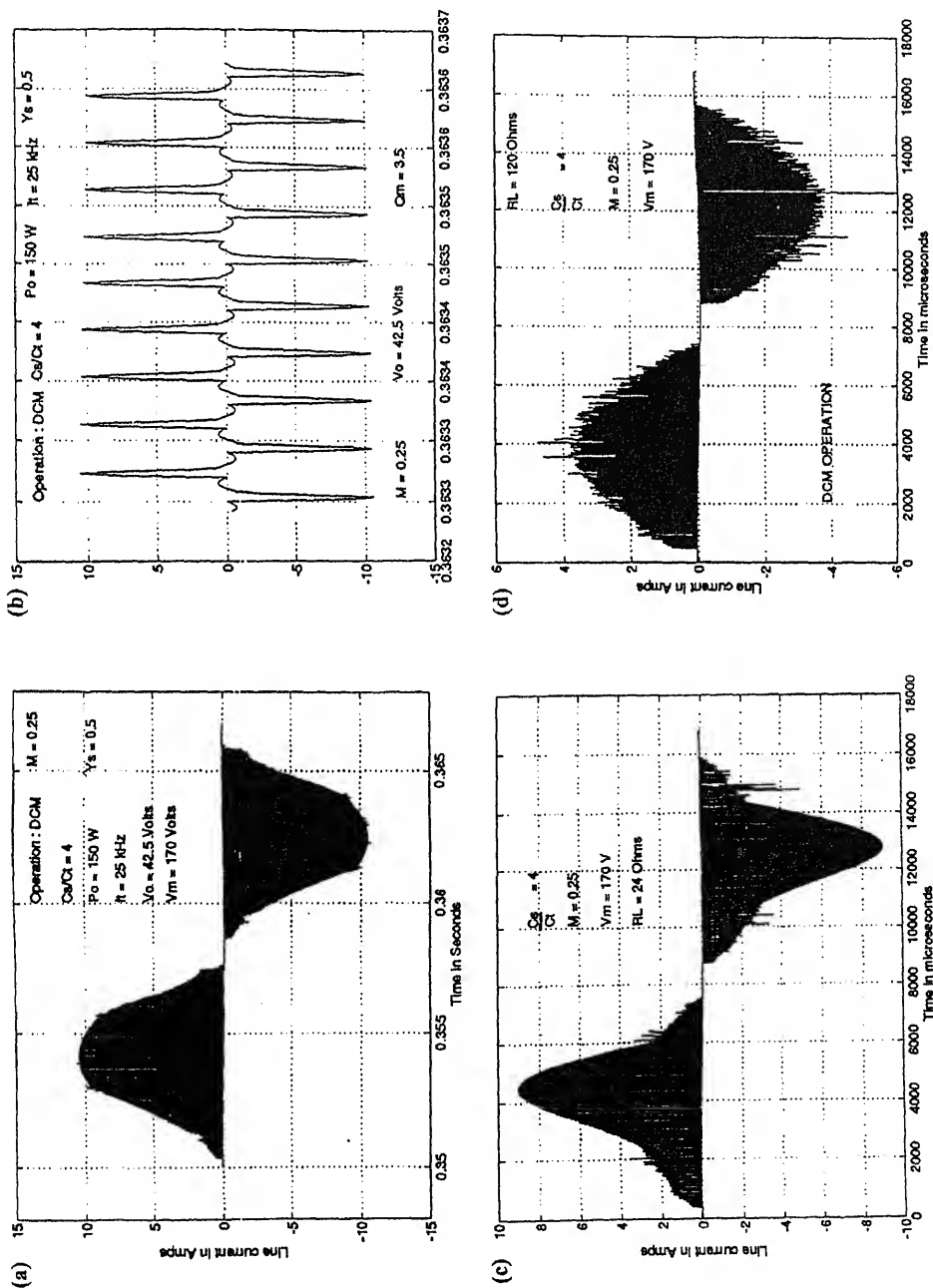
Figure 5. SPICE simulation waveforms for 150 W (full load $Q_{s\,max} = 3.5$), 42.5 V output, 25 kHz SPRC operating on the utility line ($V_{ac} = $ V and $C_s/C_t = 4$). (a) Line current i_{ac} ($P_o = 150$ W, THD = 14%). (b) Just CCM operation near the peak of the ac voltage cycle at full (HF cycles shown). (c) Line current waveform i_{ac} at 50% rated load ($P_o = 75$ W, THD = 16%). (d) Line current waveform i_{ac} at 10% rated ($P_o = 15$ W, THD = 11%).

the line current waveforms shown in figure 5 are due to insufficient HF line filtering and they can be filtered using appropriate LC filters on the ac side. It is also observed that with the choice of $C_s/C_t > 4$, the converter characteristics approaches that of the SRC.

3.4c *Experimental results:* Based on the design presented in §3.4a, an experimental prototype SPRC rated at 150 W, 42.5 V output, operating on 60 Hz, 120 V utility line was built using IRF740 MOSFET's in a bridge configuration and a 1 : 1 HF transformer. The SPRC converter was controlled using LD405 resonant controller. The filter components used were $L_d = 150 \mu H$ and $C_d = 10,000 \mu F$.

The various waveforms and line current harmonic spectra obtained from the prototype model are presented in figure 6 for different loading conditions with $C_s/C_t = 4$. In these waveforms, for reduced loads, the frequency was decreased to get the desired regulated output voltage keeping the on time constant. Figure 6a(i) shows the line voltage, line current, output filter current and output voltage waveforms, while the corresponding line current harmonic spectra at full load (THD $= 8.73\%$) are shown in figure 6a(iii). The JCCM operation of the SPRC at full load near the peak of the ac voltage cycle is shown in figure 6a(ii). For 50% rated load, line current waveform is shown in figure 6b. The THD increased to 12.9% at 50% load and then decreased to 8.9% at 10% rated load. The frequency was decreased from 68.96 kHz (at full load) to 39.8 kHz (at 10% load) in an open loop manner, to regulate the output voltage. The resonant peak current reduced from 10 A at full load to 4.5 A at 10% load. The line current waveform closely resembled sine wave with discontinuity near the valleys for entire load range. Near the valleys, discontinuity observed is due to insufficient voltage drive and converter gain to meet the required load demand and hence the distortion. All these waveforms and results closely conform with the SPICE simulations. The power factor is maintained close to unity, for the entire load range with DCM operation of the SPRC even without active control. However with active control, the line power factor can be further improved by waveshaping the line current waveform.

3.4d *Advantages and disadvantages of DCM operation:* Some of the desirable features or advantages of the described converter operating in DCM are as below.

(1) Single-stage ac-to-dc power conversion without any active control with high power factor.

(2) Since the converter operates in DCM, ZCS is assured for all the switches with negligible switching losses. Output diodes operate with ZVS. This allows very high switching frequency.

(3) The capacitor C_t is placed on the secondary-side of the HF transformer to take advantage of the leakage inductance as part of resonant inductance.

(4) The value of resonant inductance required is very small. The value of resonant capacitors are also small.

(5) The output filter inductance (L_d) is small since it has to filter twice the switching frequency.
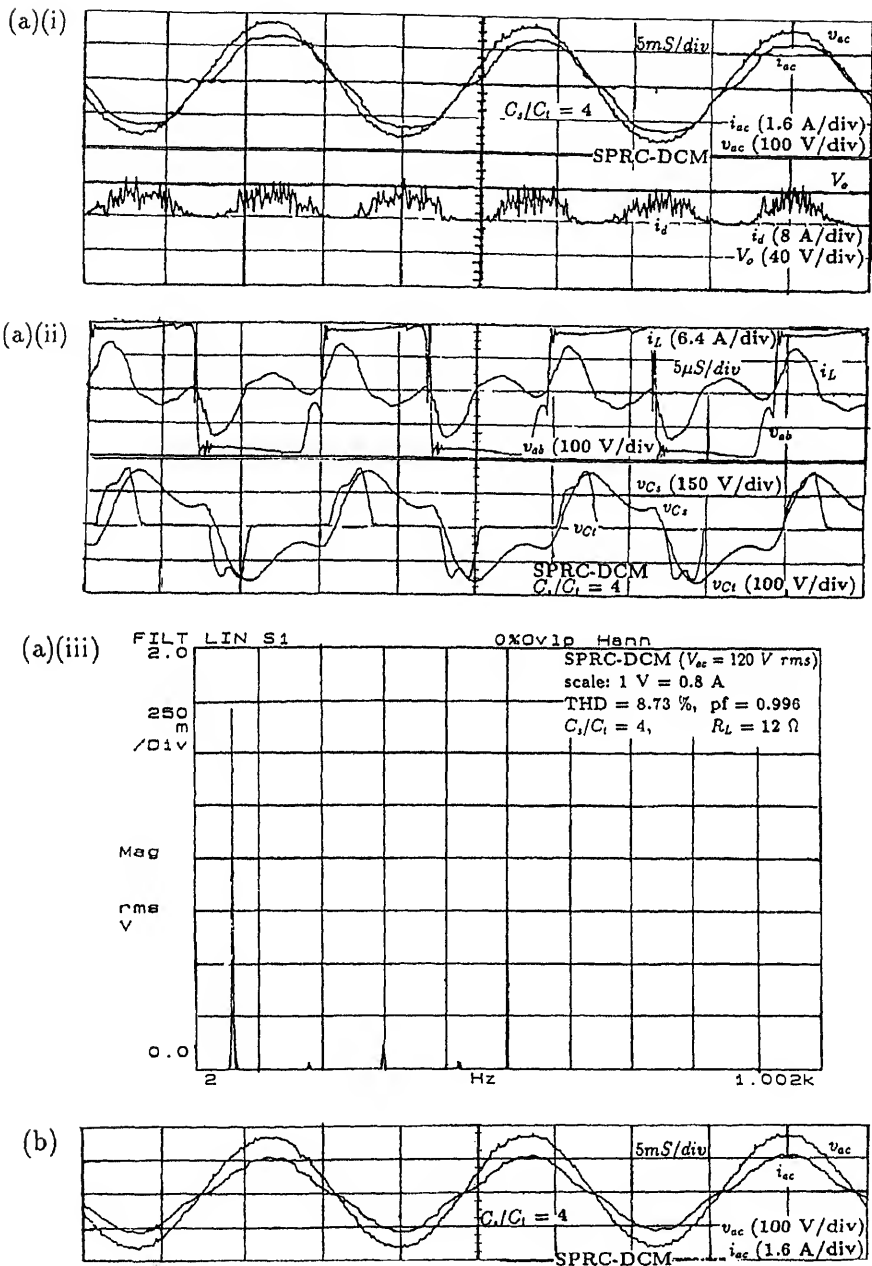
Figure 6. Experimental waveforms for the DCM converter designed in § 3.3a. **(a)** At full load ($R_L = 12\,\Omega$, $P_o = 150\,\text{W}$): (i) input voltage v_{ac}, line current i_{ac}, filter current i_d, output voltage V_o, (ii) v_{ab} and i_L, v_{Ct} and v_{Cs} near the peak of utility ac voltage, (iii) line current harmonic spectra for i_{ac}. **(b)** Half load ($R_L = 24\,\Omega$, $P_o = 75\,\text{W}$): v_{ac} and i_{ac}.

(6) In the experimental converter, MOSFETs were used as the switching devices requiring a series by-pass diode and another anti-parallel diode. IGBTs are better suited for ZCS application, also use of IGBTs is expected to reduce the conduction losses. Due to the availability of IGBTs with fast internal diodes, series diodes are not required. Mos-Controlled-Thyristors (MCTs) can also be used in the converter presented.

Some of the problems with the proposed converter are the following.

(1) Switch peak currents are quite high. Therefore IGBTs are the better switching devices, but there is a limitation on the highest switching frequency. The rectifier diode voltage ratings are high due to peak voltage across C_t.

(2) Since variable frequency operation is used for power control, lowest frequency sets the limit on the size of the HF transformer.

(3) Lossy snubbers are required.

(4) Output filter capacitor C_d has to filter the 120 Hz ripple and, therefore, is large especially for lower output voltages limiting the transient response of the converter. This problem exists in all single stage ac-to-dc converters with low harmonic distortion.

3.5 Continuous current mode of operation

The SPRC can operate in leading pf (below resonance) or lagging pf (above resonance) modes depending on the switching frequency (or pulse-width when fixed-frequency control is used), resonant component values and the load. It has been shown that operation in lagging pf (ZVS) mode has several advantages (use of lossless snubbers, no turn-on losses, use of internal diodes of MOSFETs etc.) compared to leading pf mode of operation.

Using ac complex circuit approximate analysis, it can be shown (Steigerwald 1988; Bhat 1991) that the converter gain for a dc-to-dc SPRC is given by

$$M = V_o'/V_{m,\min} = \sin(\delta/2)/[\{(\pi^2/8)(1 + (C_t/C_s)(1 - y_s^2)\}^2$$
$$+(Q_s\{y_s - (1/y_s)\})^2]^{1/2} \text{ per unit,} \tag{4}$$

where, $Q_s = (L/C_s)^{1/2}/R_L'$ and $\delta = \pi$ for variable frequency control.

Figures 7a and b show the plots of the converter gain M as a function of the normalized switching frequency y_s, for capacitance ratios of 0.5 and 1, respectively, when the converter operates as a dc-to-dc converter with constant dc input. When operated as an ac-to-dc converter as explained earlier with an HF capacitor C_i, the required variation in converter gain $M(t)$, switching frequency ratio $y_s(t)$ and $Q_s(t)$ over the 60 Hz cycle to draw nearly sinusoidal line current from the utility line when active control is used are plotted in figure 8. The sharp dip in y_s curve in figure 8 is due to insufficient gain near the zero crossings at chosen operating point. However, if no active control is used, the operating point has to be chosen carefully to satisfy all the design constraints mentioned in § 3.3. For variable frequency CCM operation, choosing y_s closer to the load-independent point on the gain curve reduces the range of variation in frequency required from full load to light load, in addition to reduction of inverter peak current stresses. Good compromised design values are: $Q_{s\,\max} = 3.2$, $M = V_o'/V_{m,\min} = 1$, $y_s = 1.153$ for $C_s/C_t = 0.5$. Similarly for
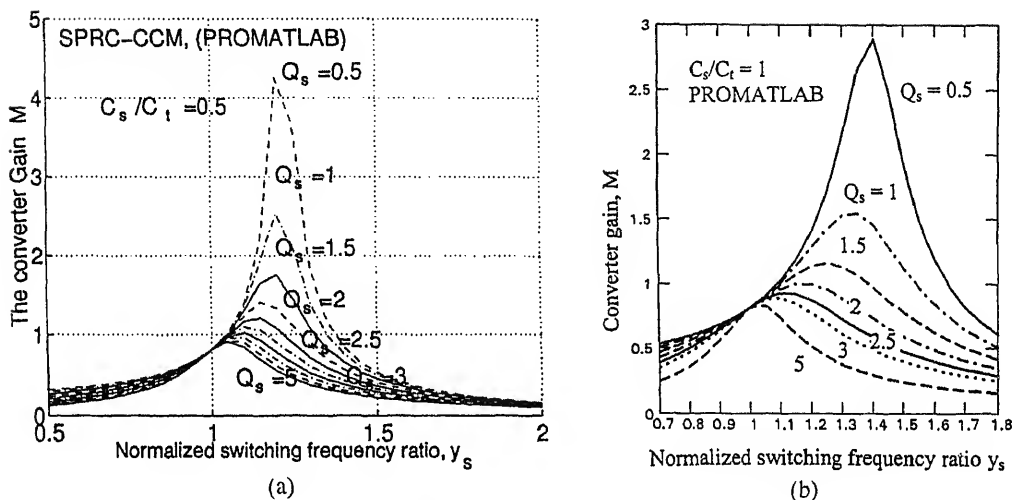
Figure 7. DC Voltage gain, M for CCM operation of the SPRC. (a) $C_s/C_t = 0.5$, (b) $C_s/C_t = 1$.

$C_s/C_t = 1$, the design values are $M = 0.75$, $y_s = 1.195$. The above design values for the converter will ensure lagging pf mode (or ZVS) operation, in addition to generating the required voltage gain near the valleys of the ac voltage, while delivering rated output power at rated minimum input voltage.

3.5a *Design example:* A converter is to be designed with the following specifications.
Average output power, $P_o = 150$ W,
Input voltage, $V_{ac} = 85$ to 110 V rms,
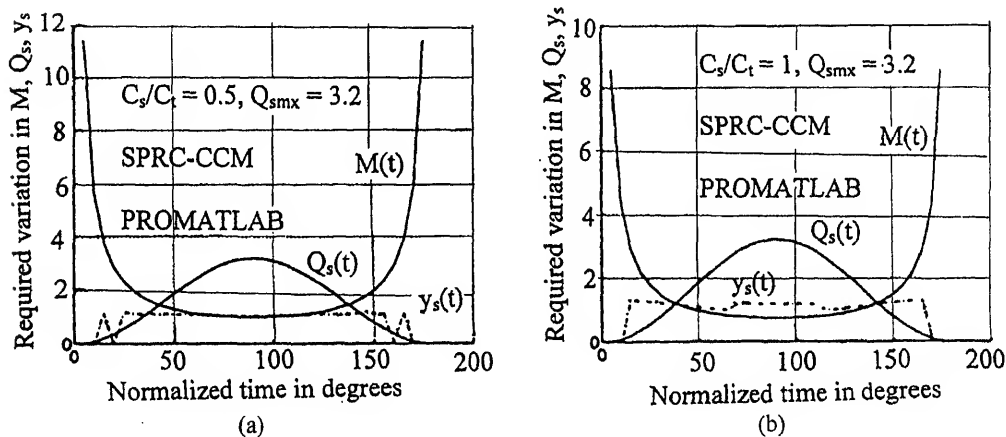
Figure 8. Variation of series $Q_s(t)$, required converter gain $M(t)$, and normalized switching frequency ratio $y_s(t)$ to get sinusoidal line current over ac half cycle for minimum ac input voltage and rated maximum load conditions for SPRC operating in CCM. (a) $C_s/C_t = 0.5$, (b) $C_s/C_t = 1$.

Output voltage, $V_o = 120\,\text{V}$,
Output current ripple, $A_i = \pm 20\%$ of I_o,
Output voltage ripple, $A_v = \pm 0.2\%$ of V_o,
Switching frequency, $f_s = 50\,\text{kHz}$.

The design calculations are done for SPRC delivering a peak power of $2P_o$ by choosing the Q_s at the peak of minimum rated line voltage. Using the relations for $Q_{s\,max}$ and y_s, the component values obtained are given below for the two cases.

Case 1: $C_s/C_t = 0.5$, $Q_{s\,max} = 3.2$, $y_s = 1.153$, $M = 1$,

$$V'o = MV_m = 120\,\text{V}, n : 1 = 1 : 1, R'_L = 96\,\Omega, L = 563.7\,\mu\text{H},$$
$$C_s = 0.0238\,\mu\text{F}, C_t = 0.0478\,\mu\text{F}.$$

Case 2: $C_s/C_t = 1$, $Q_{s\,max} = 3.2$, $y_s = 1.195$, $M = 0.75$,

$$V'_o = MV_m = 90\,\text{V}, n : 1 = 0.75 : 1, R'_L = 54\,\Omega, L = 328.65\,\mu\text{H},$$
$$C_s = 0.044\,\mu\text{F}, C_t = 0.044\,\mu\text{F}.$$

3.5b *SPICE simulation results:*   The 150 W, 120 V output, 50 kHz converter designed in §3.5a (case 1) was simulated in SPICE3 to evaluate the converter performance without active current control. For $C_s/C_t = 0.5$, the harmonic distortion in the line current waveform (figure 9a(i)) is 14.11% at full load and rated minimum input voltage. The SPRC operates in lagging pf mode near the peak, and leading pf mode near the zero crossings of the ac voltage cycle. The line current waveform for 53% load, shown figure 9a(ii) has a THD of 15.5%. The SPRC operated fully in lagging pf mode for decreased load currents due to increase in y_s for regulated output.

Similarly for a capacitance ratio of 1 and at full load, the line current THD was 19.8% (figure 9b), and the converter operated in discontinuous capacitor voltage mode (DCVM), while delivering rated output power at rated minimum input ac voltage. Since the converter operated in DCVM at full load and in CCVM at reduced load, the range of variation in switching frequency to regulate the output was larger for $C_s/C_t = 1$. In all these simulations the switching frequency f_s was increased to regulate the output voltage at reduced loads.

3.5c *Experimental results:*   Based on the design presented in § 3.5a, a breadboard model of SPRC rated at 150 W, operating on 60 Hz, 85 V to 110 V utility line was built using IRF640 MOSFET's in a bridge configuration and a readily available 1 : 1 HF transformer having 12 turns each. The SPRC was controlled using UC2825 PWM controller, configured for variable frequency operation.

(1) *Without active control* – The various waveforms obtained from the prototype model are presented in figure 10 for different loading conditions and $C_s/C_t = 0.5$. The THD obtained for the line current waveform (full load, rated minimum input voltage) presented in figure 10a(i) is 13.5%. The lagging pf and leading pf operation along with the resonant capacitor voltage waveforms at full load, near the peak and the valleys of the ac voltage
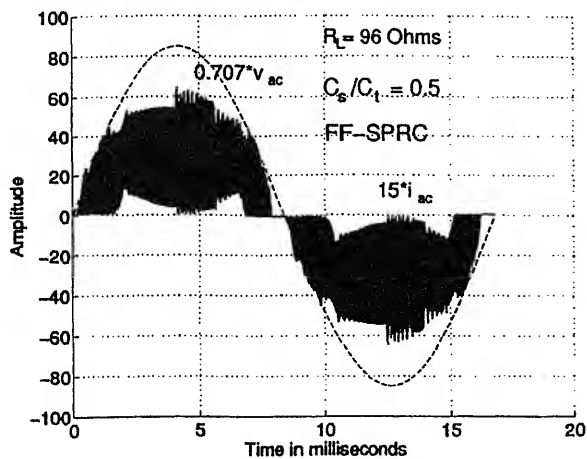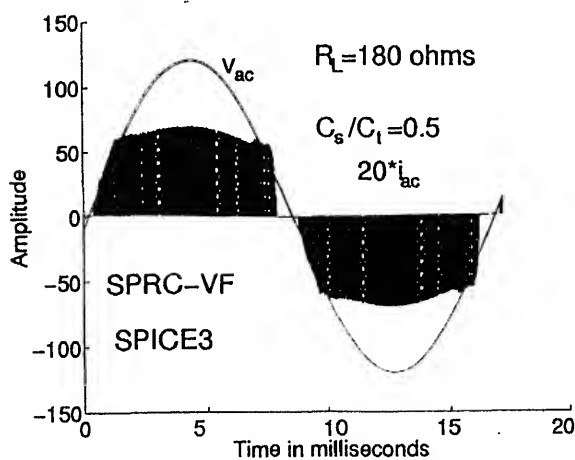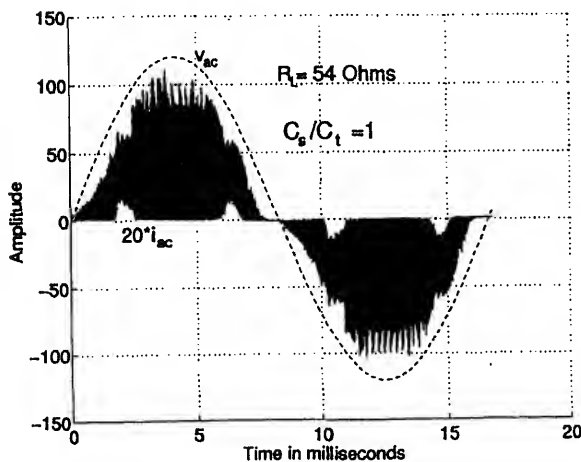
Figure 9. SPICE3 simulation waveforms for 150 W (full load), 120 V output, 50 kHz variable frequency SPRC operating on the utility line in CCM without active control ($V_{ac} = 85$ V). Waveforms v_{ac} and i_{ac} for (a) $C_s/C_t = 0.5$: (i) at full load and (ii) at 53% load; (b) $C_s/C_t = 1$: at full load.

(a)(i)

(a)(ii)

(a)(iii)

(b)
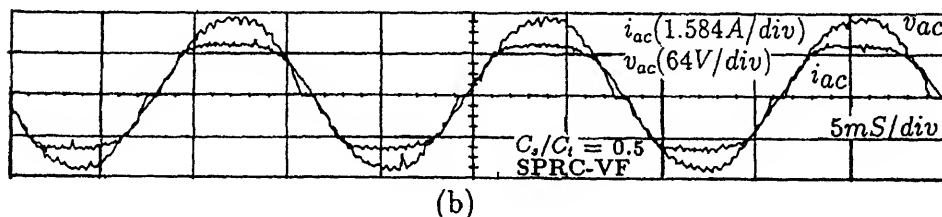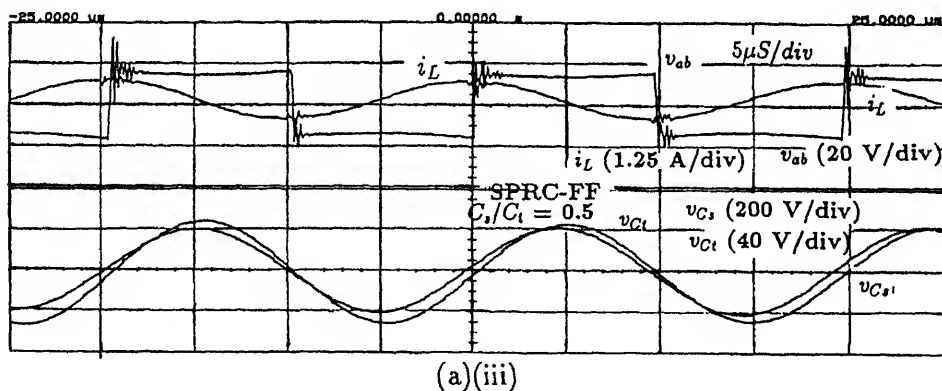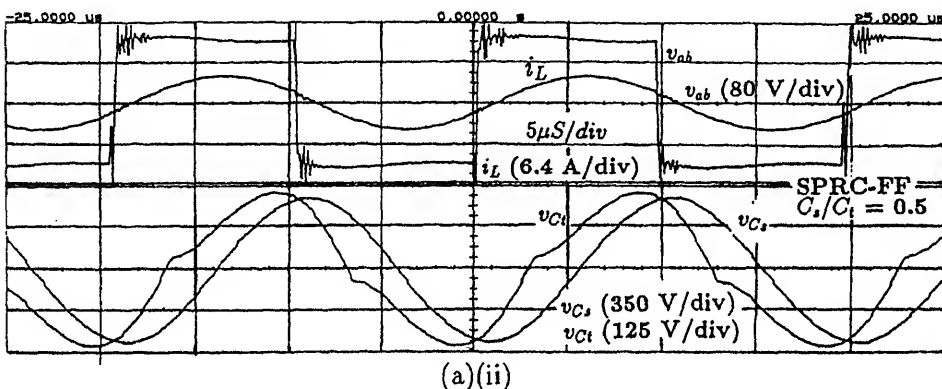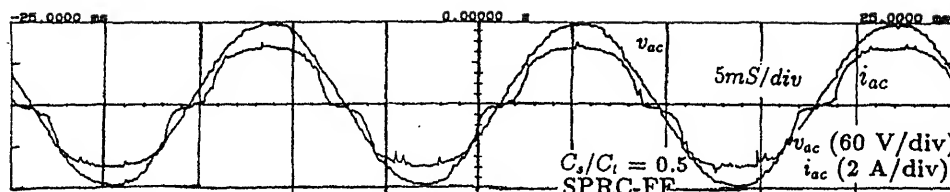
Figure 10. Experimental waveforms for different load conditions for a 120 V output, variable frequency SPRC operating on the utility line without active control ($C_s/C_t = 0.5$, $L_d = 500\,\mu H$, $C_d = 1000\,\mu F$, $V_{ac} = 85$ V rms; (a) full load ($R_L = 96\,\Omega$): (i) v_{ac} and i_{ac}, (ii) v_{ab} and i_L, v_{Ct} and v_{Cs} near the peak of ac voltage and (iii) v_{ab} and i_L, v_{Ct} and v_{Cs} near

11. Active current control scheme block diagram for SPRC bridge (figure 1) operat-
the utility line.

hown in figures 10a(ii) and (iii), respectively. At full load the predominant
omponents are 5th and 7th. The THD reached a minimum of 11.9% at 53%
line current shown in figure 10(b). The maximum distortion occurred at an
requency of 54.5 kHz, at 80% load, due to over-boosting effect at all points
cycle. The peak current reduced from 5.4 A at full load to 3.75 A at 10% load.
put voltage of 110 V rms, the THD reached a maximum (due to square wave-
minimum of 34.5% and 12.8%, at full load and 53% load, respectively. For
utput, the required variation (increase) in switching frequency for the complete
he variation is 50 kHz to 63.93 kHz. Experimental results also confirmed the
results for $C_s/C_t = 1$.

active current control – In order to reduce the line current THD, variable
ctive current control scheme has been implemented, and the details are given
block schematic of the active current control scheme implemented with the
model is shown in figure 11. In order to keep the output voltage constant for
ut voltage and output load, and also to keep the input line current close to
and in phase with the line voltage, the proposed control scheme is equipped
ntrol loops, namely:

ter voltage feedback loop: This is a slow varying loop consisting of an output
sensing amplifier, PI compensator-1 and a multiplier. The output voltage is
by a voltage divider (V_{os}) and compared with set reference signal (V_{ref}) using
mpensator. The error signal V_e in combination with sinusoidal reference $|V_{acr}|$
to generate the varying amplitude sinusoidal reference current i_r for referencing
er current control loop.

er current control loop: The current controlling feedback loop is used to monitor
ns current i_{ac} and force it to follow the mains voltage. This control loop consists
compensator-2, with inputs as, conditioned line current waveform $|i_{acr}|$ to be
and the reference current i_r. The reference current signal i_r has quick control
ine current while the output voltage error V_e has a slow control over the line
. The result is that the *dc* link current varies as a rectified sinusoid and the
regulation is achieved by adjusting the amplitude of the voltage error signal
ally the control voltage V_c to the UC2825 controller is generated by summation
i_e and V_{ea} after proper scaling and limiting circuits.
mple experimental waveforms obtained from variable frequency active control
presented in figure 12, corresponding to a capacitance ratio of 1. For rated
nput voltage, the line current waveforms presented in figures 12a(i) and b have

Figure 12. Experimental waveforms for a 90 V output variable frequency SPRC operating on the utility line with active current control (C_s/C_t = 1, L_d = 500 μH, C_d = 1000 μF, V_{ac} = 85 V rms). **(a)** At full load (R_L = 54 Ω): (i) v_{ac} and i_{ac}, (ii) v_{ab} and i_L near the peak of *ac* voltage and (iii) v_{ab} and i_L near the valleys of *ac* voltage. **(b)** 50% load (R_L = 108 Ω): v_{ac} and i_{ac}.

distortion figures of 7.4% and 11.33%, at full load and 50% load respectively. As shown in figures a(ii) and (iii) for full load, the operating frequency near the peak and valleys of *ac* voltage are 51.28 kHz and 58.82 kHz, respectively. The converter operated with ZVS operation for all the switches throughout the *ac* cycle with active control at full load.

3.6 *Fixed-frequency SPRC*

In this case, the SPRC is designed for operation with $\delta = \pi$ (equation 4) at full load with minimum rated supply voltage. Therefore, the design is the same as the variable frequency CCM case. Power control is achieved by phase shifting the gating signals to obtain a quasi-square-wave v_{AB} with pulse-width d. Based on the design example presented in § 3.5a, the converter was simulated using SPICE (Belaguli & Bhat 1995) for variable load and line voltage. It was observed that a narrow variation in pulse-width is required for regulating the output from full load to light load. It was also verified experimentally that the THD of line current is reduced by active current control scheme with pulse-width being

changed throughout the line cycle. The major problem with fixed-frequency control is that the converter operates in leading pf mode for most part of the cycle.

4. Conclusions

Operation of resonant converters on the utility line with high pf and low harmonic distortion has been presented. Design examples have been presented for the variable frequency controlled SPRC operating in DCM and CCM. The SPICE3 simulation and experimental results show that, by proper converter design, one can get low line current THD and high pf (>0.97) with SPRC even without active control. The capacitance ratio 0.5 is preferred as the THD figures, the peak current stresses and range of variation in frequency from full load to light load are lower as compared to those figures obtained for C_s/C_t ratio 1, when no active control is used. With the implementation of active control scheme, the pf is maintained close to unity (>0.99) with further reduction in THD. For active current control, capacitance ratio 1 is recommended as the THD figures are lower (8% at full load). Slightly higher current stresses obtained for C_s/C_t ratio 1 are due to DCVM operation and as a consequence require larger variation in frequency to regulate the output voltage. The experimental converter built had a switching frequency of 50 kHz at full load and this was used only to demonstrate the high pf, low THD that can be obtained with variable frequency operation of SPRC. However higher switching frequency can be used with active control scheme, as ZVS operation is maintained over the entire 60 Hz ac cycle.

References

Belaguli V 1996 *Series-parallel and parallel-series resonant converters operating on the utility line – analysis, design, simulation and experimental results*. Ph D dissertation, Department of Electrical and Computer Engineering, University of Victoria, Canada

Belaguli V, Bhat A K S 1995 Characteristics of fixed frequency series-parallel resonant converter operating on the utility line with and without active control. *IEEE Power Electronics and Drive Systems (PEDS '95)* Conference Record, pp 168–173

Bhat A K S 1991 A unified approach for the steady-state analysis of resonant converters. *IEEE Trans. on Ind. Electron.* 38: 251–259

Chambers D 1983 A new high frequency resonant technique for dynamic correction of off-line converter input current waveforms. *Proceedings of Powercon* 10, paper # F-1, pp 1–7

He J, Mohan N 1987 Input-current shaping in line-rectification by resonant converters. *IEEE Industry Applications Society Conference Records*, pp 990–995

Kataoka T, Mizumachi K, Miyairi T 1979 A pulse-width controlled AC to DC converter to improve power factor and waveform of AC line current. *IEEE Trans. Ind. Appl.* 15: 670–675

Keraluwala M H, Steigerwald R L, Gurumoorthy R 1991 A fast-response high power factor converter with a single power stage. *IEEE Power Electronics Specialists Conference Record*, pp 769–779

Kocher M J, Steigerwald R L 1983 An ac to dc converter with high quality input waveforms. *IEEE Trans. Ind. Appl.* 19: 586–599

Nijhof E B G 1986 Resonant power supply (RPS) converters: The solution for mains/line pollution problems. *Proc. Power Conversion Int. Conf. Rec.* pp 104–139

Schlecht M F, Miwa B A 1987 Active power factor correction for switching power supplies. *IEEE Trans. Power Electron.* PE-2: 273–281

Schutten M J, Steigerwald R L, Keraluwala M H 1991 Characteristics of load resonant converters in a high power factor mode. *IEEE Applied Power Electronics Conference Record*, pp 5–16

Steigerwald R L 1988 A comparison of half-bridge resonant converter topologies. *IEEE Trans. Power Electron.* 3: 174–182

Single phase power factor correction – A review

RAMESH ORUGANTI and RAMESH SRINIVASAN

Center for Power Electronics, Department of Electrical Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
e-mail: eleramsh@nus.edu.sg; ramesh.s@sp.ac.sg

Abstract. This paper provides a comprehensive review of the past and recent developments in the area of single-phase power factor correction (PFC) techniques. The motivation for the research in this area, and the manifold directions into which the research has gained impetus, are clearly brought out. The various PFC techniques are broadly classified into (1) passive, (2) active, and (3) active–passive PFC techniques. The active PFC techniques, based on the output dynamics, are further classified into (1) conventional techniques which have slow output dynamics and (2) techniques with fast output dynamics. The critical issues within each PFC technique are discussed in detail. An extensive list of references is also provided at the end.

Keywords. Power factor correction; single-phase; agency standards; IEC.

1. Introduction

Most applications requiring *ac-dc* power converters need the output *dc* voltage to be well regulated with good steady-state and transient performance. The circuitry typically favoured until recently (diode rectifier–capacitor filter) for the utility interface is cost effective, but it severely deteriorates the quality of the utility supply thereby affecting the performance of other loads connected to it besides causing other well-known problems. In order to maintain the quality of the utility supply, several national and international agencies have started imposing standards and recommendations for electronic equipment connected to the utility. Since the mid-1980's power electronics engineers have been developing new approaches for better utility interface, to meet these standards. These new circuits have been collectively called *Power factor correction (PFC)* circuits.

Reducing the input current harmonics to meet the agency standards implies improvement of power factor as well. For this reason the publications reported in this area have used "Power factor correction methods", and, "Harmonic elimination/reduction methods" almost interchangeably. Several techniques for PFC and harmonic reduction have been

reported and a few of them have gained greater acceptance over the others. Commercial IC manufacturers have introduced control ICs in the market for the more popular techniques.

In this paper, we examine and review the developments in the field of single-phase PFC techniques. Section 2 covers briefly the background information in this area. Following this, § 3 outlines a few passive PFC techniques. This is then followed by § 4 which summarises and discusses various recent developments in the field of active PFC. Section 5 then presents a few active–passive PFC techniques and their merits.

2. Background

2.1   *Conventional single phase ac-dc utility interface*

Conventionally, the utility interface of a low power single phase off-line *ac-dc* converter typically consists of a simple uncontrolled rectifier feeding a bulky filter capacitor (see figure 1a). An input π filter, not shown in the figure, is usually present in the *ac* side in commercial products to reduce the electromagnetic interference (EMI) due to the converter. The bulk capacitor, C_{in}, is designed to maintain the ripple in the *dc* bus to an acceptable level and also to meet the "hold-up time" requirements. The circuit draws narrow input current pulses around the line voltage peaks (see figure 1b).

2.1a   *Power factor:*   Power factor (PF), is defined as the ratio of the average power to the apparent power,

$$PF = \frac{\text{average power}}{\text{apparent power}} = \frac{V_s I_{sl} \cos \theta_1}{V_s I_s} = \frac{I_{s1}}{I_s} \cos \theta_1. \tag{1}$$

Power factor reflects how effectively the given source power is utilised by the load. In (1), it has been assumed that the input voltage is sinusoidal with low distortion. Here, V_s is the *rms* input voltage, I_s is the *rms* input current, I_{s1} is the *rms* value of fundamental input current, and θ_1 is the phase angle of the fundamental current.

(a)   (b)
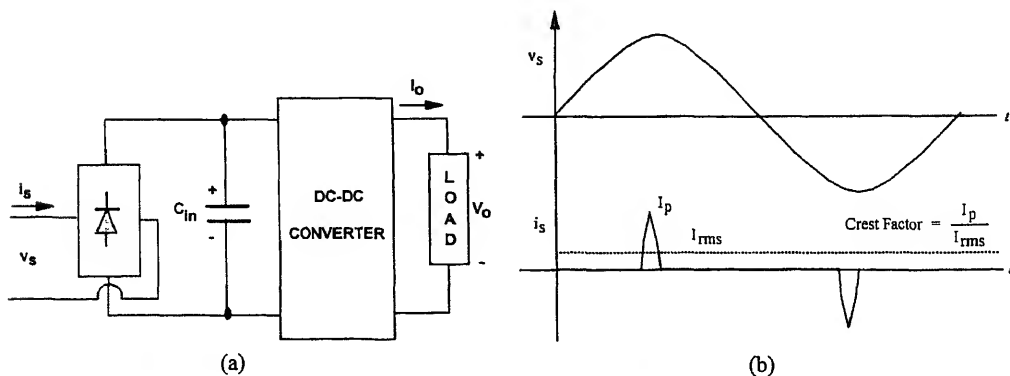
Table 1. Typical input current harmonics of the single-phase rectifier–capacitor type interface.

n	3	5	7	9	11	13	15	17
$(I_{sn}/I_{sl})\%$	87	80	78	70	63	50	41	30

Note: Even harmonics are zero in magnitude

The ratio I_{s1}/I_s is the *distortion factor*; its value reflects the effect due to input current harmonics. The cosine of the angle θ_1 is the *displacement power factor*; its value reflects the conventional power factor when there is no distortion.

1b *PF and harmonic components of a conventional interface:* The narrow current pulses drawn by the conventional rectifier–capacitor type interface is rich in harmonics. The typical harmonic amplitudes of such an interface is listed in table 1 as a percentage of the fundamental[4]. The third, fifth, seventh, ninth and the eleventh harmonics, are considerable in magnitude.

The displacement factor being close to unity, the PF in a conventional interface is strongly linked to the distortion factor. The PF of operation is also quite poor (typically 0.60).

2 Utility issues and agency standards

2a *Problems of conventional interface:* The large harmonic content (see table 1) and the consequent poor PF of operation of the conventional rectifier–capacitor type interface causes several problems to the utility supply. Some of them are listed below. References [1–4] discuss these in greater detail.

) *Due to harmonic components* – Because of the non-zero source impedance in the utility supply, the harmonic currents flowing through it will cause the voltage waveform to be distorted at the point of common coupling to other loads. This may cause malfunction of other loads and also of power system protection and metering devices. Besides voltage waveform distortion, harmonic components can also cause the following problems.

) Overheating of the neutral line.

) Interference with communication and control signals.

) Overvoltages due to resonance conditions.

) Overheating of the distribution transformer and distribution lines.

i) *Due to poor PF* – Poor power factor of operation implies ineffective use of the volt–ampere ratings of the utility equipment such as transformers, distribution lines and generators. Also, it places a restriction on the total equipment load that can be connected a typical home or office wall-plug with specified maximum *rms* current rating

2.2b *Agency standards:* Due to the proliferation of power electronic equipment connected to the utility system, the concern over the pollution of the utility supply has been growing stronger over the years[7–10]. It has been reported that 5% of the utility generated power in USA is consumed by desk-top computers alone [5]. In order to facilitate the delivery of quality power to end-users and also to utilise the existing power generation and distribution equipment more effectively, several national and international agencies have started imposing strict standards that specify the extent of harmonic pollution that can be tolerated in the line current. One such standard (International Electrotechnical Commission IEC-555-2 or EN-1000-3-2) is expected to be implemented in Europe shortly [6]. Countries like USA and Japan are also showing strong inclination to impose similar standards.

2.3 Desirable features of a PFC technique

Input side features:

(1) Sinusoidal input current with close to unity PF operation.

(2) Reduced EMI.

(3) Insensitive to small signal perturbations in the load (i.e., good output-to-input susceptibility figure).

Output side features:

(1) Good line and load regulation.

(2) Low output voltage ripple.

(3) Fast output dynamics (i.e., high bandwidth).

(4) Multiple output voltage. levels if needed by the application.

Others: Electrical

(1) Galvanic isolation between input and output.

(2) High power conversion efficiency.

(3) Hold-up time if required.

(4) Universal input voltage operation (85 V–270 V *ac rms*).

Mechanical

(1) Low part count.

(2) Smaller size and weight.

Economical/environmental

(1) Low cost.

(2) Power save feature (green supply).
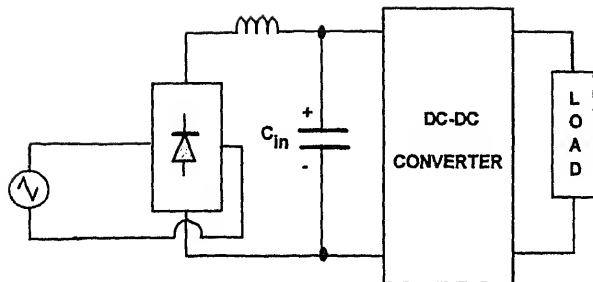
Figure 2. Conventional rectifier circuit with inductive filter.

3. Passive PFC techniques

The power line disturbances caused by the proliferation of phase controlled and diode rectifier circuits were of concern even in late 70s [3,11]. The definition of power factor for nonlinear circuits and passive techniques for improving it are presented in an early literature [11]. Currently, passive techniques remain attractive for low power PFC applications [16,19]. It has been reported [19] that power factor as high as 0.98 can be achieved using passive PFC techniques. The following sub-sections discuss a few of the passive PFC arrangements.

3.1 *Inductive filter*

Figure 2 shows a well-known scheme [4,11], with an inductor inserted between the output of the rectifier and the capacitor. The inclusion of the inductor results in larger conduction angle of the current pulse and reduced peak and *rms* values.

For low values of inductance the input current is discontinuous and pulsating. Typical PF achieved in discontinuous mode operation (DCM), with practical values for the inductor, is in the range of 0.65 to 0.75. Better power factor (PF) is achievable by using a larger value of the inductance and pushing the operation to continuous conduction mode (CCM). However, it is shown [11] that even for infinite value of the inductance, the PF cannot exceed 0.9 for this arrangement.

The inductor may also be introduced on the *ac* side [4]. The position of the inductance will not affect the PF in DCM operation. Under CCM operation, however, the circuit behaviour itself will be different depending upon the location of the inductance. For instance, the presence of an infinite inductance on the *ac* side (i.e. CCM operation) will result in zero input current and zero voltage across the bulk capacitor, theoretically. However, the same inductance on the *dc* side will result in rectangular blocks of current in the input with the bulk capacitor charging up to the average value of the rectified input voltage.

For lower power levels, the distribution line inductance will itself act as a good filter. For an office plug point (15 A), the line inductance is typically in the range of 1 to 4 mH. An estimate of this value can be obtained from the assumption that the *ac* side reactance X_s $(= \omega L_s)$ is 5% of the ratio of nominal rated voltage to the maximum current rating of the plug point [4]. A practical value for the line impedance seen at a particular wall-plug outlet, may also be obtained using the method suggested [72].
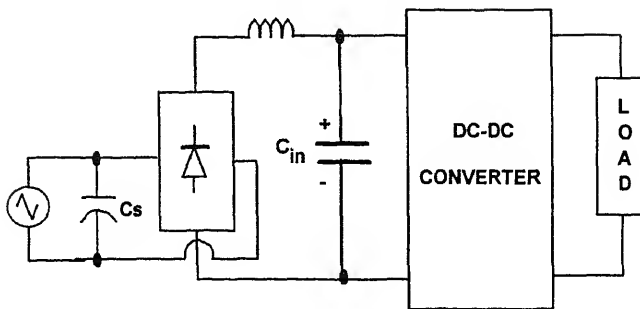
Figure 3.   Rectifier circuit with input capacitor C_s.

In the scheme shown in figure 3, a small filter capacitor C_s is connected across the input terminals of the circuit. The line inductance (not shown in figure 3) and C_s forms the first stage LC filter. Therefore higher order harmonics of the line frequency will undergo greater attenuation (typically 80 dB) resulting in better harmonic performance. It is reported [12] that even for relatively small values of the inductance, a PF of 0.86 is attainable, which is a considerable improvement over the no-capacitance case.

3.2   *Resonant input filter*

Figure 4 shows the series filter arrangement for power factor correction [13–14], which results in good power factors as high as 0.94. Thus, harmonic performance is also good. However, the power factor depends upon the resonant quality factor which is load dependent. Here the bandpass filter is designed with a centre frequency equal to the supply frequency. The quality factor "Q" determines the bandwidth and hence the harmonic content of the supply current. High "Q" (narrow bandwidth) will result in reduced harmonic content and close to unity power factor. This circuit arrangement is popularly used in applications where the supply frequency is high. One such application is the space platform [13–14], with supply frequency up to 25 kHz.

Some authors [15,17], suggest the use of parallel resonant filter (see figure 5) for PF improvement. With this arrangement power factor close to 0.95 is achieved. The filter is tuned to offer a very high impedance to the third harmonic component (the most predominant). The high value parallel resistor is added to damp out circuit oscillations.
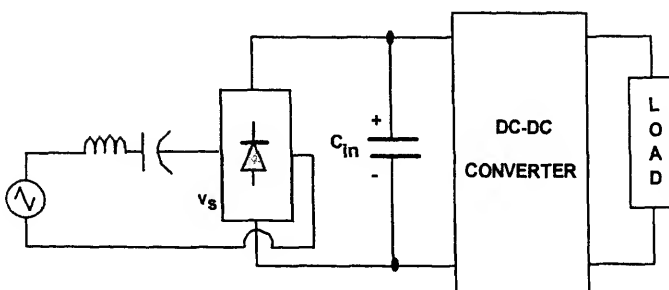
Figure 4.   Rectifier circuit with series resonant filter.
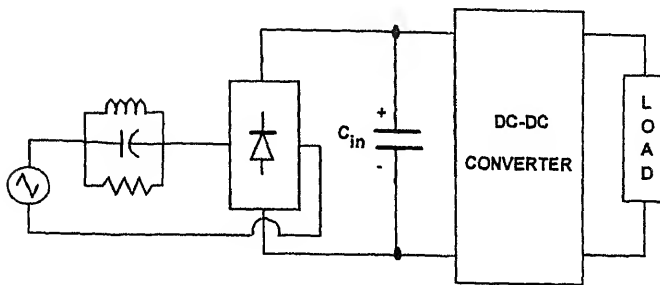
Figure 5.   Conventional rectifier circuit with parallel resonant filter.

3.3   *Evaluation of passive techniques*

Passive filters may prove to be a good solution [16–18] for switch mode power supplies (SMPS) operating at lower power levels (<200 W). Typical applications are TVs, VCRs, other home entertainment electronics, and perhaps some of the lower powered office automation equipment such as computers, FAX machines etc.. It is to be noted that in these applications the real aim of the filter is to contain the harmonics within the agency specified limits [6] and not improvement of power factor *per se* [93]. At such power levels the passive solution is economical [16,19], and more efficient and may also result in size benefits. However, even for low power levels, there are a few factors that make the passive solution less attractive than the active solution in certain cases. The factors are as below.

Universal input voltage range (85 V to 280 V ac, 47 to 63 Hz):   It is desirable for both the manufacturers and the users that the equipment works in all these supply conditions without any modifications. This requirement is difficult to meet using passive PFC technique whereas it is an inherent feature in many active PFC techniques.

Optimal use of the wall outlet power: Passive PFC technique results in lower PF than active PFC technique. Therefore, passive PFC technique is less attractive when priority is given to optimising the use of the wall plug's volt-ampere capacity rather than just meeting the agency standards. Optimising the wall plug's VA capacity allows a user to plug more equipment to the same wall outlet.

   For higher power levels, however, the passive solution suffers from disadvantages such as:
(1) Large size of the reactive elements [17,24].
(2) May not be able to contain the harmonics within the agency specified limits (the agency requirements are different for different power levels [6]).
(3) Poor power factor compared with active schemes.
(4) Not cost effective.
   However, the passive solutions are simple to understand and implement, besides being

The active PFC technique, which involves the shaping of the line current using switching devices such as MOSFETs (metal oxide semiconductor field effect transistors) and IGBTs (insulated gate bipolar junction transistors) is a result of advances in power semiconductor devices and microelectronics. For low and medium power ranges up to a few kilowatts (<5 kW), MOSFETs are by far the popular choice for PFC because of their switching speed, ease of driving and ruggedness. BJTs and more recently IGBTs are used for high voltage medium power applications which MOSFETs are unable to contend with owing to their large on-state resistances.

For achieving good input current waveshaping using active techniques, typically the switching frequency should be at least an order of magnitude greater than 3 kHz (= 50 × 60 Hz = 50th harmonic of line frequency). With modern advances in MOSFETs and IGBTs, this is feasible.

The use of active PFC techniques results in one or more of the following advantages.

- Lower harmonic content in the input current compared to the passive techniques.
- Reduced *rms* current rating of the output filter capacitor.
- Near unity power factor (0.99) is possible to achieve with the Total Harmonic Distortion (THD) as low as 3–5%.
- For higher power levels active PFC techniques will result in size, weight and cost benefits over passive PFC techniques.

The following sub-sections present the recent advances in single phase active PFC techniques. The active PFC techniques have been classified into two broad categories in this survey.

(1) Active PFC techniques with poor load dynamics. These have been referred to in this paper as "*Conventional active PFC techniques*" (§ 4.1). They are typically followed by a *dc-dc* downstream converter which caters to the demands of the load.
(2) Active PFC techniques with fast load dynamics (§ 4.2). Here, the PFC unit is capable of meeting the fast dynamic requirement of a typical load.

4.1 *Conventional active PFC techniques*

Here, there are basically two approaches. One approach is to use current-source-type circuit [47–51], in which the PFC acts as a current source feeding the load. Using the other approach results in the well known voltage-source-type circuits discussed [20–23,25–46,52–54, 57–68,70–72,75–95,102]. Though the voltage source type circuits are more popular, the current source type circuits are useful in certain niche applications. In the following subsections both types of circuits and schemes are discussed.

4.1a *Topologies:* The current-source-type PFC converter [47–51] is usually of buck type. However, for the voltage-source-type PFC converter, any one of the basic *dc* to *dc* converter switching cells, such as buck, boost, buck-boost and Cuk converter, can be used. Among
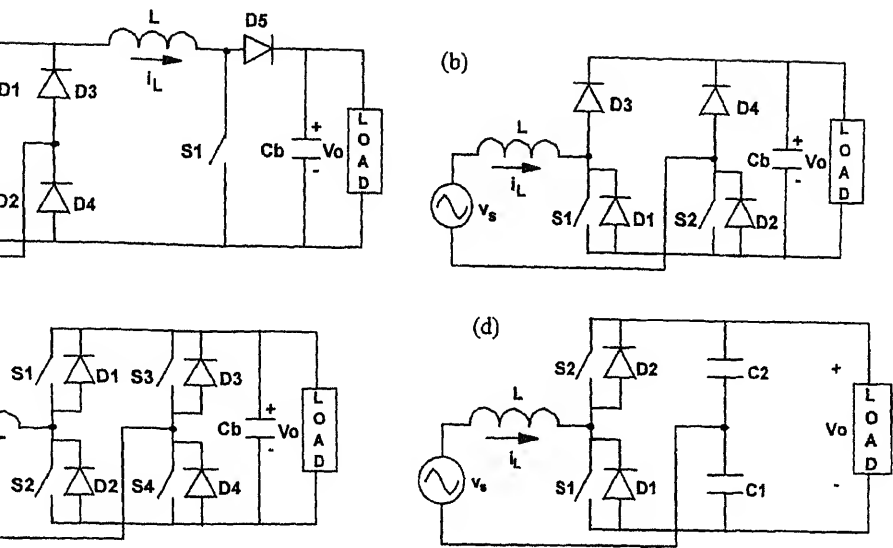
Figure 6. Circuit variations of the boost PFC topology. Single-switch (**a**), two-switch (**b**), four-switch (**c**), and half-bridge (**d**) boost PFC's.

boost and the buck-boost topologies are more popular. In the following sub-
the notable features of these topologies (both current and voltage source types)
on critical issues related to these topologies are presented.

topology – In figure 6, some of the popular boost circuits are shown. Several
the boost converter such as the position of the inductor on the input side (reduced
current ripple and EMI), and a *dc* voltage gain greater than unity, make it a natural
active PFC application [4]. The single-ended boost converter (figure 6a) is more
due to its use of a single active switch and also due to the ease of driving the switch.
modes of operation are possible for the single-ended boost converter topology.

tinuous conduction mode (CCM) [25].

dary of CCM and DCM [27].

ntinuous conduction mode (DCM) [102].

M: This is a very popular choice for medium and high power applications.
e supply ripple current due to switching is low because of the presence of the
st input inductor. Hence, the input filtering requirement is relatively low. Isolated
boost PFC converter is also available [29]. The following issues are to be noted
to the boost converter operating in CCM.

ecovery loss: The reverse recovery loss of the output diode is a serious problem
gle-ended boost converter operating in CCM [70–71]. The output diode should
everse blocking voltage equal to the output voltage, which for the universal input

This reverse recovery of the diode also increases the turn-on losses in the boost switch. This factor places a limitation on the switching frequency. Hence, the size of the unit cannot be reduced beyond a certain limit. The reverse recovery also causes increased EMI [71].

Charge dumping loss: Another factor that limits the switching frequency is the parasitic output capacitor (typically around 150 pF) of the boost switch, usually a power MOSFET. The capacitor's energy is lost in the channel of the switch during turn-on; the resulting power loss could be significant at high switching frequencies.

EMI performance: In a conventional off-line converter (figure 1), the line current is narrow and peaky. However, the waveform is still smooth with very little switching component owing to the presence of a bulky capacitor, C_{in}, at the output of the bridge rectifier feeding the *dc-dc* converter (see figure 1a). As a result, the input EMI filter size can be small. In the case of boost PFC converter, however, the switching frequency components are directly fed into the supply lines. Hence, a large EMI filter is required in spite of the presence of the boost inductor, L. The issues related to EMI filtering are presented [73,74].

Cusp distortion: The single-ended boost converter (figure 6a) operating in CCM suffers from line current distortion near the zero crossings [43,54]. This is because the reference current slope at zero crossings is higher than the slope of the charging current of the boost inductor which defines the maximum current rate at which the input current can be increased. The effect of this "cusp" distortion is to introduce line current harmonics. In DCM operation, cusp distortion may be lower, as only a small value of inductor will be used.

One or more of the foregoing problems may be addressed by using loss-less switching techniques. By turning the switch on during the instants when the voltage across it is zero (zero voltage switching – ZVS) the problem of loss due to charge dumping is solved. ZVS also helps to reduce the EMI. In converters using resonant techniques [58–59] to realise ZVS, even though switching losses are minimised, conduction losses and switch voltage stresses are higher compared to those using PWM control. This problem, however, is less severe in converters using PWM soft switching techniques [103–105]. Hua *et al* [105], a soft switching technique for a boost PFC circuit.

Boost in CCM – DCM boundary: The converter operating at the boundary of CCM and DCM is also a popular PFC technique used mainly for low power applications such a electronic ballasts for lighting [27]. The operation of the converter in CCM–DCM boundary eliminates the problems due to reverse recovery of the output diode. However, the charge dumping loss is not prevented. The high value of the switching currents places a severe stress on the switch, diodes and the output capacitor. The input filtering requirements are also high. It is this factor that limits its application to low power levels.

Boost DCM: As a front-end PFC converter in a cascaded scheme (§ 4.3a), the boost circuit operating in DCM mode is seldom used, as it offers no special advantage over that
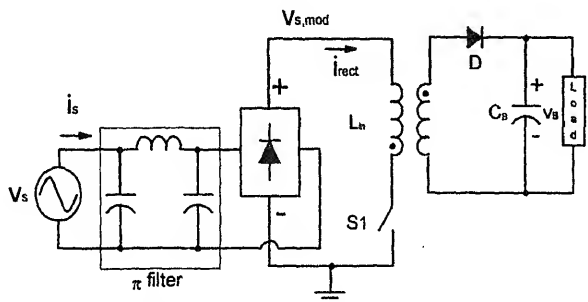
Figure 7. Flyback PFC converter.

erating in the boundary of CCM and DCM. However, operation in DCM is popular single-stage PFC implementations (see § 4.3c) that use boost circuit as front part of converter. Unlike the flyback converter discussed in the following paragraphs, a boost nverter operating in DCM does not yield a sinusoidal input current (neglecting the itching components) if the duty ratio is held constant over a line cycle.

Buck-boost or flyback topology – The flyback topology (figure 7) is attractive for low wer applications.

The start-up inrush current problem faced by the boost topology is not present here.

The implementation of the overload protection is simple compared with the boost topology.

The output voltage may be greater or less than the peak input voltage.

Implementation of galvanic isolation between the input and the output is simple with flyback topology.

Both CCM and DCM modes are possible for flyback PFC converter [26,36]. But in either de the input current is chopped resulting in more noise and EMI than with a compa- le boost topology. However, the DCM operation eliminates the diode reverse recovery blem.

The DCM operation of the flyback is popular because of the simplicity of control. usoidal input current (after filtering the switching component) is drawn automatically he switch duty cycle is maintained constant over one line half cycle [26].

Watson *et al* [103,104] used a flyback topology as shown in figure 8. Here, the switch S2 ith its anti-parallel diode) together with clamp capacitor C_c forms an active-snubber
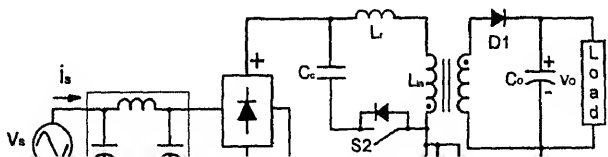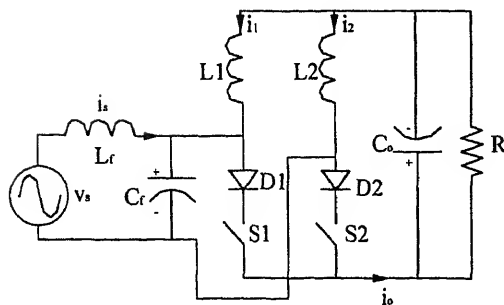
Figure 9. Power circuit of the half-bridge buck converter.

network. Due to this, the energy stored in the transformer leakage inductance L_{in} is not dissipated; instead it is recycled while minimising turn-off voltage stress across the power switch S1. Furthermore, the ZVS of switch S1 lowers output rectifier di/dt resulting in reduced rectifier switching loss and output switching noise.

(iii)   Buck topology – In voltage source type buck circuits, the input voltage must be greater than the output voltage. Due to this, during certain intervals when the input voltage is less than the output voltage, no current will be drawn from the input; hence, the net input current will be nonsinusoidal. But the advantage of the buck converter is that it can provide current limit support due to the series buck switch [102].

In the application reported in references [38–40], the input voltage varies over a very wide range (130 to 600 V) but the output voltage desired is 400 V. Thus, the boost converter alone is not suitable. So, a topology with (Buck + Boost) operation is adopted to maintain the input current sinusoidal. When the input voltage is greater than 400 V, the buck operation is effective. Below 400 V, the boost operation is performed.

Unlike the voltage-source-type buck converter, the current-source-type buck PFC converter operates for all values of the input voltage drawing a sinusoidal current. To achieve this, however, the instantaneous output current has to be maintained at a higher value than the instantaneous input current.

A three-phase buck converter operating as a current-source-type PFC converter is also reported [47,48]. The application of the idea to a single-phase case is also mentioned [47,48]. References [49–51] deal with the single-phase implementation itself. Reference [51] discusses a two-switch current source type PFC circuit (figure 9) which has been derived from a half-bridge voltage source type PFC converter [45,46] based on the duality principle. Note that, in these converters, a diode must be added in series with the switch(es) to realise bipolar blocking capability; this results in additional conduction power loss. The current-source-type buck PFC converter is particularly useful in application where the output voltage must be reduced to low values, such as in a dc motor control or ac-dc rectifier with over current limit.

4.1b   *Input current control techniques used in conventional PFC converters:*   (i) *Peak current mode control* – In this constant frequency control method, the switch (peak) current is sensed and made to track the desired sinusoidal reference current [55,56,65]. The current reference is obtained by multiplying the error signal of the output voltage control loop with

e sinusoidal template obtained from the input voltage. The peak current mode control easy to implement and has an inherent fast current limit protection. Several integrated cuits such as ML4812 [55], have been developed to implement this control.

The peak current mode control technique is not suitable for the buck and the flyback C topologies due to the large error between the switch peak current and the average ut current which is to be controlled. The control method is however suitable for a boost nverter, though there will be a "peak to average" error in this topology [56] also.

The peak current mode control technique suffers from duty cycle dependent instabil- ("subharmonic oscillation"). In a *dc-dc* converter, this problem can be eliminated by oviding "external ramp compensation" [56]. In case of an *ac-dc* PFC converter, external mp compensation is not very effective as the operating point, which is line and load pendent, varies over a wide range. Furthermore, providing excess ramp compensation ads to input current distortion, particularly around zero-crossings of the input voltage.

) *Average current mode control* – By controlling the average input current instead of the ak [20,54,55,57], the subharmonic oscillation problem is avoided. Good input current wave aping is achieved as the average line current is directly controlled. Due to the averaging ter used, the dynamic performance of the current control loop is not fast, which however not seen as a major limitation in PFC applications. Several commercial integrated circuits ch as UC3854 [54,57] are available for the implementation of this control for boost PFC nverter operating in CCM.

i) *Charge control* – In charge control, the average input current is controlled on a cycle- -cycle basis giving rise to fast current loop response. Though this technique is general, is particularly useful for flyback PFC[35,36] where the use of average current control may t give the desired performance. As with peak current mode control, the charge control chnique is also reported to exhibit subharmonic oscillation [35].

) *Hysteresis current control (HCC)* – In this control method, the input current is made switch within a reference current window called the hysteresis band [30,32,34,45,46,94]. ere are many variations in this technique such as constant hysteresis band, and variable steresis band [32]. The variable frequency operation of this method is seen as one of its sadvantages as the energy storage elements including input filter may have to be sized r the lowest frequency of operation. HCC technique, however, offers several advantages ch as ease of implementation, and fast and robust current control.

) *Sinusoidal pulse width modulation control* – Here, just as in inverter applications, a angular carrier-wave is modulated with a sinusoidal modulating signal to generate the ive pulses for the switch(es) [28]. Closed loop regulation is achieved by varying the depth modulation.

i) *Delta modulation control (DMC)* – This technique [69], which is the dual of the HCC ethod, is usually used with current-source-type *ac-dc* PFC converter [47,48] (see figure 9). ere, the capacitor (C_f) voltage is made to switch within a predetermined band about the ference sinusoidal voltage. For a given sinusoidal input voltage, the input current then

automatically becomes a sinusoid. The power factor of the input current can be adjusted to any value by appropriately adjusting the phase of the reference voltage. However, DMC is parameter sensitive and relatively complex leading to the proposal of the next control technique [50,51].

(vii)   Inductor voltage control (IVC) – In this technique [50,51], the input inductor (L_f) voltage (figure 9) and hence the input current is controlled to follow a sinusoidal reference voltage, resulting in a significant performance improvement over the DMT, such as robustness, simplicity and direct input current control.

4.1c   *Modelling of PFC converters: (i)   Small signal modelling of conventional PFC circuits* – In most PFC applications, the output voltage of the PFC is dynamically regulated against load and line variations, using closed loop control. In order to design a good compensator circuit based on analysis, linear small-signal models are necessary for the power and control circuits. These models are dependent on the topology and the control technique used.

Mohan *et al* [20] present the small signal model for the boost converter operating in CCM with the load assumed to be resistive. However, in most cases, the load of a conventional PFC would be of constant power type with negative impedance characteristics. The small signal model for the boost converter with this type of load is also discussed [25,57,67,68].

(ii)   Large signal modelling of PFC converters – The large signal "averaged" model can be used to obtain several steady-state results of the system easily. Besides steady-state analysis, the averaged model is very useful in efficient simulation (using SPICE, SABER etc.) of complex systems such as the PFC circuit. Both simulation time and the memory storage requirement for the output are significantly less than in the simulation using actual switches. Due to these, unlike with a switched model, the averaged model simulation can be easily run several times to optimise converter and controller performance. There are a variety of approaches to obtain an averaged model. These are dealt with in detail in other references [96–101]. The averaged model concept for PFC application is used extensively [46,51,93,94].

The "switched" model with ideal switches, despite its drawbacks such as long simulation time and large secondary storage space, is still useful and necessary in the study of switching transients and in the determination of quantities such as peak voltage and current stress of the device, duty ratio and switching frequency variations. Thus, the averaged and the switched models complement each other in their usefulness.

4.2   *Problems in conventional active PFC techniques*

All conventional PFC techniques discussed in § 4.1 suffer from some drawbacks which are discussed in this section. These form the motivation for the "Active PFC techniques with fast load dynamics" discussed in § 4.3.
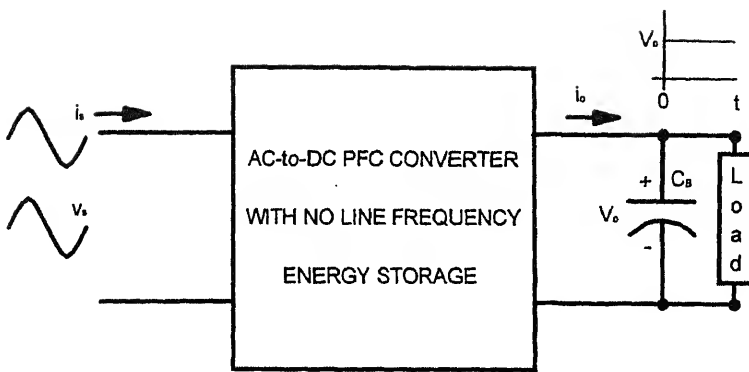
Figure 10. Input–output properties of a PFC converter.

us input and output power must be equal. With sinusoidal input current and voltage, instantaneous input power is a sine-squared function (figure 11) consisting of a *dc* ver component plus an *ac* component with frequency at twice the line frequency. At the put port, the output voltage is regulated to a constant value and hence the load draws stant (*dc*) power only. Due to this, there must be a bulk energy storage element (such as in figure 10) to absorb the second harmonic input power. The amplitude of the second monic power is the same as the *dc* value delivered to the load, hence requiring a large pacitor.

The second harmonic energy can also be stored in an inductor such as in current-source-e PFC converters. However, with the present technology, the size and cost of a capacitor generally much lower than a corresponding inductor. This is the main reason for the esent popularity of voltage-source PFC converters over current-source PFC converters.

b   Slow output dynamics: A typical closed loop control system for a PFC is shown figure 12. In order to have fast dynamic response for sudden changes in line voltage l load, the bandwidth of the loop must be high. But, as the magnitude of the second monic voltage component at the output is quite high (even with large C_B), the control nal V_c (figure 12) will then have a large second harmonic component. However, in ler to achieve a sinusoidal input current it is essential that V_c be slow varying and be
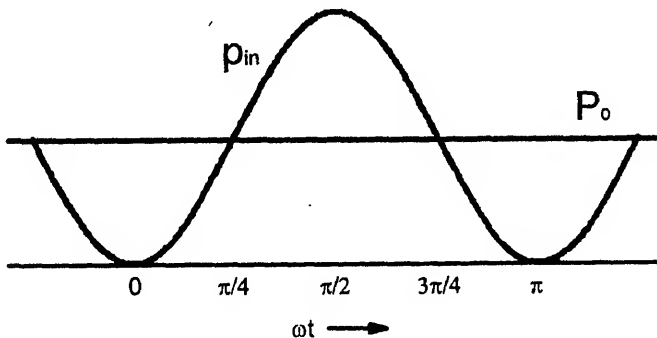
Figure 11.   Input and output power waveforms.
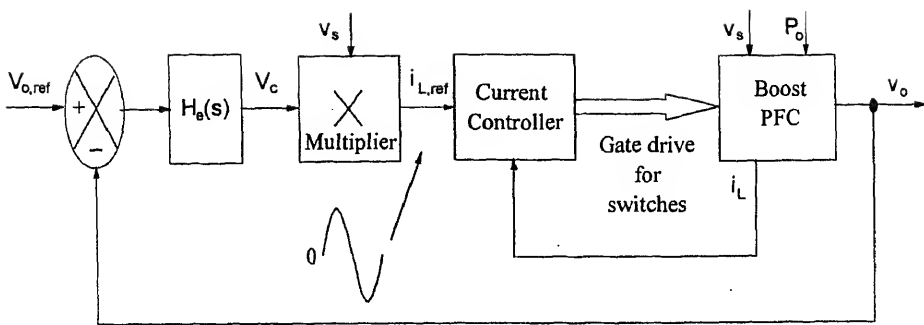
Figure 12. Block schematic of a typical closed loop system for PFC converter.

maintained constant at least over one half of the line cycle. Due to this design conflict, the requirement of fast dynamics is usually sacrificed in order to achieve good input current waveform. The typical bandwidth of the outer voltage loop will be of the order of 5–20 Hz, resulting in slow output dynamics.

Note: The response for line voltage variations, however, is usually good due to a input voltage feed-forward that is typically adopted [102].

4.3 Active PFC techniques with fast load dynamics

As pointed out in § 4.2, due to the existence of the second harmonic problem in conventional active PFC techniques, the output voltage-control-loop will have a poor response. Approaches to overcome this limitation are discussed in the following sections.

4.3a *Cascaded scheme:* This scheme [102] is currently very popular. Here, the conventional PFC is cascaded with a down stream *dc-dc* converter (see figure 13). The downstream converter addresses the load requirements such as providing fast dynamics, tight regulation etc., and the PFC converter meets the line requirements, thus ensuring independent control over the input and output requirements. Also, the resulting PFC arrangement is modular in construction. Several ICs are available, ML 4819 [55] for example, with the PFC and DC/DC converter control functions integrated. With all its advantages, the cascaded scheme still suffers from the following drawbacks.
(1) The rated output power is handled twice resulting in poor efficiency.

(2) Having separate full power rated modules for the PFC and the downstream *dc-dc* converter entails higher cost, weight and size.

Figure 13. Power flow in a cascaded scheme.
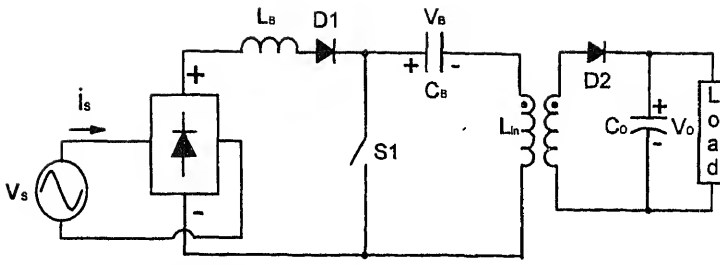
Figure 14. A single-stage PFC scheme – Boost integrated with flyback rectifier/energy storage dc-dc converter (BIFRED).

3b *Second-harmonic filtering scheme:* The dynamics of conventional PFC converters can be improved by filtering-out the second harmonic component of line frequency from entering the error amplifier stage of the control loop by using a notch filter [75,76]. The error amplifier can now be designed to have a wide bandwidth and high-gain for the low frequency range resulting in fast response. However, with the use of this technique, no corrective action will be taken by the loop for perturbations at around twice the line frequency. It must be noted that, this technique does not prevent the second harmonic of the line frequency from reaching the load; it only avoids it from reaching the control loop.

3c *Single-stage PFC schemes (S^2PFC):* A variety of single-stage schemes featuring tight output regulation and near-sinusoidal input current with power factor close to unity have been proposed [78,89]. Most single stage schemes, such as the BIFRED[83,86], BIBRED [83], DITHER [79] are realised by combining the two power stages of a cascaded scheme into a single power stage allowing the active switch(es) to be shared (figure 14). Because of the sharing of the switches, the control freedom is greatly reduced. Also, the voltage and current stresses of the active switches are typically much higher than those of the switches in the cascaded scheme, thereby resulting in less efficiency. Often variable frequency PWM techniques [79,82,83] or resonant techniques [80] are employed in S^2PFC scheme.

In many variable frequency PWM techniques, two control variables (say switch on-time and frequency) are still used for meeting the line and the load requirements independently. Jovan *et al* [78], however, report a single-stage scheme operating with fixed frequency PWM
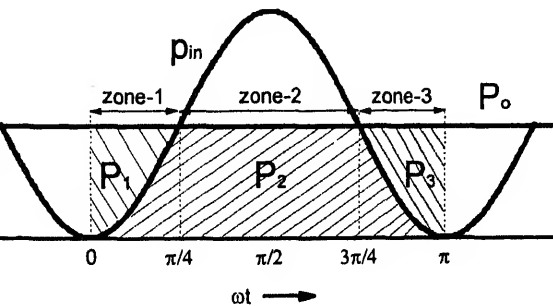
Figure 15. Input and output power waveforms to illustrate P^2PFC concept.
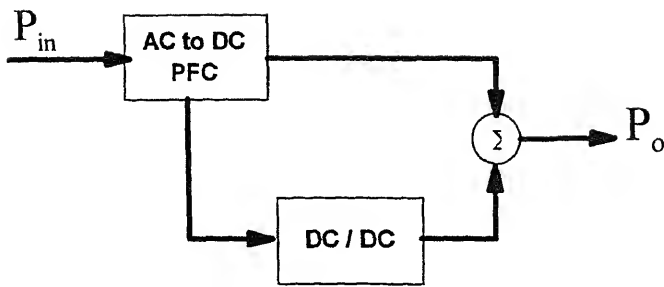
Figure 16. Power flow in a P²PFC scheme-1.

technique. Here, a two-switch forward converter is integrated with the boost PFC. One of the switches is shared by both the converters.

Generally, single-stage schemes suffer from one or more of the following drawbacks.

(1) Suffers from all the drawbacks of DCM operation of conventional PFCs.

(2) In many cases, the voltage across the bulk energy storage capacitor is uncontrolled and can reach high values. A higher rated capacitor will result in increased cost and greater power loss due to larger ESR (Equivalent Series Resistance) values.

(3) The frequency varies in many cases over a wide range (typically 1 : 8) making the EMI filter design difficult.

(4) Increased stress on devices.

(5) Efficiency of power conversion is generally low.

(6) Complex control.

4.3d *Partially parallel PFC schemes (P²PFC):* In figure 15, the input and output power waveforms are divided into three (time) zones. In zone-2, the input power processed by the PFC is greater than the output power demand. In a conventional PFC scheme, this excess energy flows into the output bulk capacitor or inductor. In the P²PFC scheme, however, this excess power is diverted to a bulk capacitor (or a bulk inductor) within the system.

During zone-1 and zone-3, the input power processed by the PFC is less than the output power demand. Therefore, in order to meet the load power requirement, the excess energy stored previously in zone-2 is now released to load (through another power conversion). The sum of energy deficits during zones 1 & 3 is equal in value to the excess energy stored within the system during zone-2.

It can be shown that the average value of the excess power during zone-2 is equal to 32% of the output power P_o. Thus, in this scheme, 68% of the output power is directly fed to the load after one-time processing by the PFC stage. The remaining 32% is however
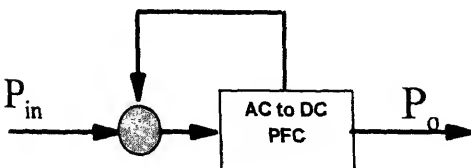
Figure 17. Power flow in a feed back type of P²PFC scheme-2.
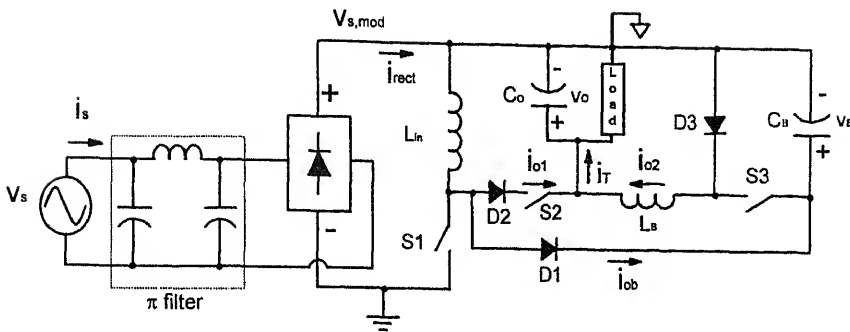
Figure 18. A PFC circuit based on P²PFC scheme-1.

ocessed twice before it is fed to the load. As a result, the overall efficiency of such a heme can be expected to be higher than that of the cascaded scheme. The challenge, re, lies in designing an appropriate control scheme that ensures fast output dynamics hile enabling the input current to be a sinusoid with PF close to unity.

Figures 16 & 17 show two possible power flow structures that can be used to implement e P²PFC concept. Several circuit implementations are possible to realise each of the wer flow structures. In the circuit (figure 18) proposed [106], the *ac-dc* PFC stage is nplemented using a flyback converter operating in DCM. The *dc-dc* converter stage is nplemented using a buck converter. The flyback converter delivers 68% of the output wer directly to the load. The remaining 32% is handled once by the flyback PFC stage id for the second time by the *dc-dc* converter stage before it is fed to the load. A few her circuit implementations conforming to the P²PFC power flow structure are also ported [77,90]. Fast output dynamics is achieved by appropriately controlling the PFC age and the *dc-dc* converter stage.

In figure 19, a circuit based on flyback topology that implements P²PFC scheme-2 is own [77,91]. Here, the PFC stage feeds the rated power to the load, but part of the output wer (32%) is fed back to itself via C_B to be processed again during zones 1 and 3.

3e *Output shunt PFC scheme (OSPFC):* This scheme is referred to [92] as Reactive unt Regulator Scheme and is shown in figures 20 and 21. Here, the second harmonic wer processed by the PFC stage is diverted to the *dc-dc* converter stage and the *dc* power
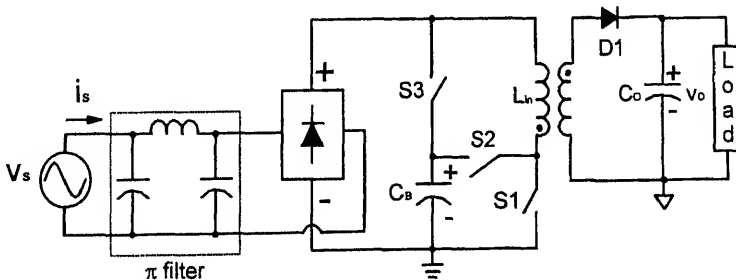
Figure 19. A PFC circuit based on P²PFC scheme-2.
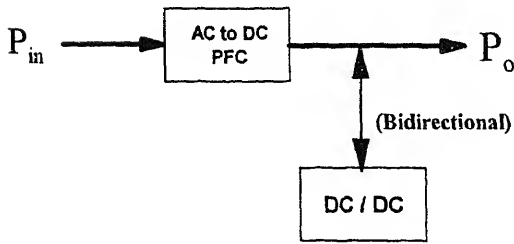
Figure 20.   Power flow in OSPFC scheme.

is fed to the load. It is easy to see that in this arrangement, the *dc-dc* stage processes 64% (i.e. $2 \times 32\%$) of P_o while the PFC stage processes full power P_o. Hence, the improvement in efficiency over the cascaded scheme may not be significant.

4.3f   *Input-shunt PFC schemes (ISPFC):*   Unlike the OSPFC scheme, the shunt converter here is an *ac-dc* bi-directional converter and it is located at the *ac* input side. The literature reports two approaches for the ISPFC scheme which are discussed in the following paragraphs.

(i)   Scheme-1 – This scheme (figure 22) has been in use for several years in large power systems for harmonic elimination [60–64]. However, this can be successfully used for correcting the power factor of small systems also[64]. The off-line converter in the main power flow path draws the well-known nonsinusoidal current (see figure 1b). In order for the net input current to be sinusoidal, the *ac-dc* shunt PFC stage is made to draw a current which is the difference of the desired sinusoidal input current and the off-line converter's input current. It must be noted that the PFC stage does not process any active power; it only handles the harmonic power. Because of the high value of the peak input current drawn by the off-line converters, the *ac-dc* PFC stage should be designed to handle high current stress.

(ii)   Scheme-2 – This scheme [94,106], consists of a cascaded scheme paralleled at the input with an off-line converter (figure 23). With this arrangement, the net input current is
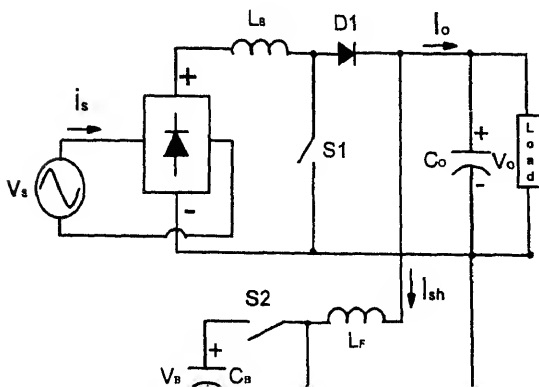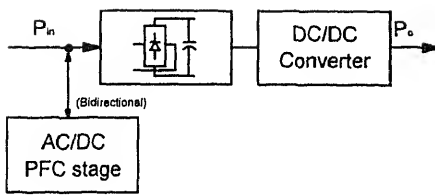
Figure 22.   Input-shunt PFC scheme-1

controlled in such a way that it is sinusoidal and in phase with the input voltage. Therefore, (as in scheme-1) the current drawn by the PFC stage will be the difference of the sinusoidal input current and the narrow and peaky current drawn by the rectifier–capacitor front-end. Again, as in scheme-1, the *ac-dc* PFC converter here handles bi-directional power flow. However, as it also processes active power, the peak current stress here is relatively low [94,106]. For a certain range of power sharing between the two parallel paths, this scheme may result in a better efficiency than the cascaded scheme [94,106].

At the output, the *dc-dc* converters #1 and #2 may each be supplying independent loads (figure 23a), and may each have multiple output voltage levels. However, for the case when there is only one load, the outputs of the two *dc-dc* converters may be paralleled to feed the load (figure 23b). For ensuring stable parallel operation, current-mode control technique may be used in the *dc-dc* converters. The sharing of the load power may also be programmed using this control technique.

4.3g   *Fully parallel PFC scheme (FPPFC):*   A fully parallel PFC scheme is proposed [93] in which a PFC stage is paralleled, both at the input and the output, with a conventional off-line *ac-dc* converter with rectifier–capacitor front-end (figure 24). The power flow is shared between *dc-dc* converter and *ac-dc* PFC stage. Thus, the overall power is processed once only. Due to this arrangement, the net input current is nonsinusoidal; though the PFC stage draws a sinusoidal current, the off-line converter draws the well-known narrow and spiky currents. The main aim in this scheme is to only contain the line current harmonics within the agency specified limits rather than achieving a high power factor. At the output, fast dynamics is achieved by using the off-line converter to cancel the second harmonic component injected by the PFC stage. Here again, it is shown that a 68% power sharing by the PFC stage and 32% power sharing by the off-line converter results in good performance. This technique may prove to be economical for lower power levels.

(a)

(b)
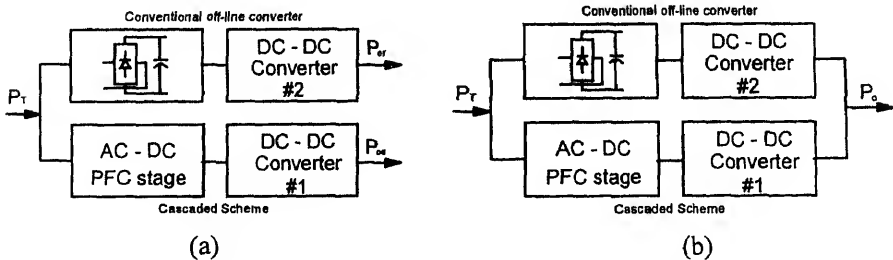
Figure 23.   Input-shunt PFC scheme-2 (ISPFC-2). (a) Independent loads. (b) Outputs paralleled.
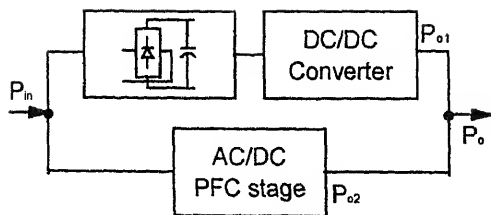
Figure 24. Block diagram of a fully parallel scheme (FPPFC).

5. Active–passive PFC techniques

As the name indicates, active–passive PFC involves the combination of active and passive techniques for PFC/harmonic reduction. The main motivation for this technique is improvement of efficiency (over the cascaded scheme) and possible size and cost reduction.

The input currents of the active–passive schemes reported so far[70–71], are not pure sinusoids. The attained PF though low (0.75 to 0.9) is, however, significantly higher than the conventional off-line converter with rectifier–capacitor type interface. Active–passive scheme can be used for applications where the main aim is to contain the line current harmonics within the agency specified limits rather than achieving a high power factor. This technique may prove to be economical for lower power levels.

In the scheme by Elmore et al [70], a conventional boost PFC is provided with a bypass passive path consisting of the boost inductor, boost output diode and the output capacitor. The boost switch is not operated around the line voltage peaks; during this interval, the passive path conducts. A simple open loop control is adopted [70], which results in large output voltage variations and uncontrolled power sharing between the two parallel paths.

Oruganti & Thean [71] suggest a loop control scheme, in which both the output voltage and the power sharing are directly controlled. The control scheme, however, is somewhat complex. The merits and demerits [71] of the active passive scheme is given below.

(1) The diode reverse recovery loss problem, occurring in a boost converter, which is maximum around the line voltage peaks is now reduced. Hence, efficiency is improved. Or else, for a given efficiency, the switching frequency can be pushed higher to achieve cost and size reduction.

(2) Reduced EMI as the line current wave form does not have any switching current component around the peaks.

(3) Input current is non-sinusoidal and hence higher harmonic content. Power factor is also poor when compared with the conventional PFC.

(4) Does not alleviate the problem of slow output dynamics. Thus the active–passive PFC stage must be further cascaded with a dc-dc converter.

(5) Not suitable for universal input voltage range as the output voltage varies with ac input voltage magnitude.

6. Conclusions

As seen in the paper, the research activities in the area of single-phase PFC can be identified as falling under three categories: (1) Passive PFC technique, (2) active PFC technique, and

(3) active–passive PFC technique. Passive PFC techniques result in significant improvement in PF and harmonic performance compared to the rectifier–capacitor type interface. However, in general, these are not superior compared to what is achievable using active PFC techniques. It is stated in many publications that passive PFC technique does not offer cost, size, and weight benefits. However, such a statement is very general. This is because, if the main aim is to contain the line current harmonics within the agency specified limits and not improvement of PF *per se*, then the passive PFC technique may prove to be a better choice for lower power levels (<200 W). A designer must investigate this aspect further for his/her specific application.

The work in active PFC technique was classified in two ways, in this paper: (1) Conventional active PFC (poor output dynamics); (2) Active PFC with fast output dynamics.

The conventional PFC circuits interface well with the utility supply in terms of high PF (near unity) and reduced harmonics. However, they cannot be used to feed loads that require fast output dynamics. Several alternative PFC schemes to overcome this problem, such as cascaded, S^2PFC, P^2PFC, OSPFC, ISPFC, and FPPFC have been reported. In all these the primary aim of the researchers has been to achieve one or more of advantages such as (1) improved performance, (2) higher efficiency and (3) cost, size and weight benefits.

The work reported in active–passive PFC schemes is not extensive. Schemes have been reported in which an active PFC path is paralleled with a passive rectifier to realise several advantages such as increased efficiency, reduced EMI, and increased switching frequency.

References

[1] J S Subjak, J S McQuilkin 1990 Harmonics – Causes, effects, measurements, and analysis: An update. *IEEE Trans. Ind. Appl.* IA-26: 1034–1042
[2] R P Stratford 1980 Rectifier harmonics in power systems. *IEEE Trans. Ind. Appl.* Vol. IA-16: 271–276
[3] W Shepherd, P Zand 1979 *Energy flow and power factor in nonsinusoidal circuits* (London: Cambridge University Press)
[4] N Mohan, T M Undeland, W P Robbins 1994 *Power electronics: Converters, applications, and design* (New York: John Wiley & Sons)
[5] Solid state. *IEEE Spectrum*, January 1994, p. 50
[6] International Electrotechnical Commission Sub-Committee 77A 1992 Disturbances in supply systems caused by household appliances and similar electrical equipment. Draft Revision of IEC Publication 555.2
[7] T S Key, Jih-Sheng Lai 1993 Comparison of standards and power supply design options for limiting harmonic distortion in power systems. *IEEE Trans. Ind. Appl.* 29: 688–695
[8] J Klein, M K Nalbant 1990 Power factor correction – Incentives, standards and techniques. *Power Conversion Intell. Motion* (June): 26–30
[9] D Maliniak 1992 DC supply keeps line power factor near unity. *Electron. Design* (May 28): 33–34
[10] K Umezu 1993 AC power supply system measures harmonics from electronic equipment

[12] S B Dewan 1981 Optimum input and output filters for single phase rectifier power supply. *IEEE Trans. Ind. Appl.* IA-17: 282–288

[13] Vorpe'rian, R Ridley 1990 A simple scheme for unity power-factor rectification for high frequency AC buses. *IEEE Trans. Power Electron.* 5: pp 77–87

[14] P Jain 1994 A unity power factor resonant AC/DC converter for high frequency space power distribution system. *IEEE Power Electron. Spec. Conf. Rec.* 1:

[15] A R Prasad, P D Ziogas, S Manias 1990 A novel passive waveshaping method for single phase diode rectifiers. *IEEE Trans. Ind. Electron.* IE-37: 521–530

[16] K Yamashita 1994 Harmonics fighters pursue choke-coil, one converter power supplies. *Nikkei Electron. Asia* (August): 44–47

[17] Jih-Sheng Lai, D Hurst, T Key 1991 Switch mode power supply power factor improvement via harmonic elimination methods. *IEEE Appl. Power Electron. Conf., Rec.*, pp 415–422

[18] A R Prasad, P D Ziogas, S Manias 1988 A comparative evaluation of SMR converters with and without active input waveshaping. *IEEE Trans. Ind. Electron.* IE-35(3): (August) 461–468

[19] R Redl, L Balogh 1995 Power-factor correction in bridge and voltage-doubler rectifier circuits with inductors and capacitors. *IEEE Appl. Power Electron. Conf., Rec.*, pp 466–472

[20] N Mohan, T M Undeland, R J Ferraro 1984 Sinusoidal line current rectification with a 100 kHz B-SIT step-up converter. *IEEE Power Electron. Spec. Conf., Rec.* : 92–98

[21] C P Henze, N Mohan 1986 A digitally controlled *ac-dc* power conditioner that draws sinusoidal input current. *IEEE Power Electron. Spec. Conf., Rec.* : 531–540

[22] M F Schlecht, B A Miwa 1987 Active power factor correction for switching power supplies. *IEEE Trans. Power Electron.* PE-2: 273–281

[23] M F Schlecht 1986 Harmonic-free utility/DC power conditioning interfaces. *IEEE Trans. Power Electron.* PE-1: 231–239

[24] R Oruganti, Y T Cham 1993 Power factor correction techniques in off-line switching power supplies. *J. Inst. Eng. Singapore* 33: 27–35

[25] L H Dixon 1990 High power factor switching preregulator design optimization. Unitrode Power Supply Design Seminar Manual, SEM 700, pp 13–1 to 13–12

[26] J LoCascio, M Nalabant 1990 Active power factor correction using a flyback topology. *Power Conversion Intell. Motion, Conf., Rec.*

[27] B Andreycak, C H Yeam, J A O'Connor 1991 UC3852 controlled on-time zero current switched power factor correction preregulator. Design Guide (Preliminary), Application note U-132, Unitrode Power Supply Design Seminar Manual, SEM 800

[28] P N Enjeti, R Martinez 1993 A high performance single phase AC to DC rectifier with input power factor correction. *IEEE Appl. Power Electron. Conf., Rec.*, pp 190–195

[29] E X Yang, Y M Jiang, G C Hua, F C Lee 1993 Isolated Boost Circuit for Power Factor Correction. *IEEE Appl. Power Electron. Conf., Rec.*, pp 196–203

[30] M S Dawande, G K Dubey 1995 Bang bang current control with predecided switching frequency for switch mode rectifiers. *IEEE Power Electron. Drives Syst. Conf., Rec.*, pp 538–542

[31] M S Dawande, G K Dubey 1993 Programmable input power factor correction method for switch mode rectifiers. *IEEE Appl. Power Electron. Conf., Rec.*

[32] C Zhou, R B Ridley, F C Lee 1990 Design and analysis of hysteretic boost power factor correction circuit. *IEEE Power Electron. Spec. Conf., Rec.*

[33] C Zhou, M M Jovanovic' 1992 Design trade-offs in continuous current-mode controlled boost power factor-correction circuits. *Proceedings of the High Freq. Power Conversion*

[34] J J Spangler, A K Behera 1993 A comparison between hysteretic and fixed frequency boost converters used for power factor correction. *IEEE Appl. Power Electron. Conf., Rec.*, pp 281–286

[35] W Tang, F C Lee, R B Ridley, I Cohen 1992 Charge control: Modelling, analysis and design. *IEEE Power Electron. Spec. Conf., Rec.*

[36] W Tang, Y M Jiang, G C Hua, F C Lee, I Cohen 1993 Power factor correction with flyback converter employing charge control. *IEEE Appl. Power Electron. Conf., Rec.*, pp 293–298

[37] H Endo, T Yamashita, T Sugiura 1992 A high power factor buck converter. *IEEE Power Electron. Spec. Conf., Rec.*

[38] Y M Jiang, F C Lee 1994 A new control scheme for buck + boost power factor correction circuit. Volume V of the Virginia Power Electronic Center publication series

[39] M C Ghanem, K Al-Haddad, G Roy 1993 A new single-phase buck-boost converter with unity power factor. *IEEE Ind. Appl. Soc. Annu. Meeting, Rec.*, pp 785–792

[40] M C Ghanem, K Al-Haddad, G Roy 1996 A new control strategy to achieve sinusoidal line current in a cascade buck-boost converter. *IEEE Trans. Ind. Electron.* 43: 441–449

[41] D Maksimovic, R Erickson 1995 Universal-input, high-power-factor, boost doubler rectifiers. *IEEE Appl. Power Electron. Conf., Rec.*, pp 459–465

[42] A H Mitwalli, S B Leeb, G C Verghese, V J Thottuvelil 1996 An adaptive digital controller for a unity power factor converter. *IEEE Trans. Power Electron.* 11(2): 374–382

[43] J C Salomon 1993 Techniques for minimising the input current distortion of current-controlled single-phase boost rectifiers. *IEEE Trans. Power Electron.* 8: 509–520

[44] J C Salomon 1993 Circuit topologies for single-phase voltage-doubler boost rectifiers. *IEEE Trans. Power Electron.* 8: 521–529

[45] J T Boys, A W Green 1989 Current-forced single phase reversible rectifier. *Inst. Elec. Eng. Proc.* B136: 205–211

[46] R Srinivasan, R Oruganti 1997 Analysis and design of single phase power factor correction with half bridge boost topology. *IEEE Appl. Power Electron. Conf., Rec.*, 1: 489–499 (Also sent to *IEEE Trans. Power Electron.*)

[47] Bakari M M Mwinyiwiwa, P M Birks, Boon-Teck Ooi 1992 Delta-modulated buck-type PWM converter. *IEEE Trans. Ind. Appl.* 28: 552–557

[48] Boon-Teck Ooi, Bakari M M Mwinyiwiwa, X Wang, G Joos 1991 Operating limits of the current-regulated delta-modulated current-source PWM rectifier. *IEEE Trans. Ind. Appl.* 38: 268–274

[49] S Funabiki, N Toita, A Mechi 1991 A single-phase PWM AC to DC converter with a step up/down voltage and sinusoidal source current. *IEEE Ind. Appl. Soc., Annu. Mtg., Rec.*, pp 1017–1022

[50] R Oruganti, M Palaniapan 1996 Inductor voltage controlled variable power factor buck type *ac-dc* converter. *IEEE Power Electron. Spec. Conf., Rec.*, pp 230–237

[51] R Srinivasan, M Palaniapan, R Oruganti 1997 A single phase two-switch buck type AC-DC converter topology with inductor voltage control. *IEEE Power Electron. Spec. Conf., Rec.* pp 556–563 (Also sent to *IEEE Trans. Power Electron.*)

[52] M J Kocher, R L Steigerwald 1983 An AC-to-DC converter with high quality input waveforms. *IEEE Trans. Ind. Appl.* IA-19: 586–599

[53] L Balogh, R Redl 1993 Power-factor correction with interleaved boost converters in continuous-inductor-current mode. *IEEE Appl. Power Electron. Conf., Rec.*, pp 168–174

[54] P C Todd 1993–94 UC3854 controlled power factor correction circuit design. *Unitrode – Product Applications Handbook*, Lexington, MA

[55] *Micro Linear Data Book* 1993

[56] *Unitrode Applications Handbook* 1987–88 Lexington, MA

[57] L Dixon 1991 Average current mode control of switching power supplies. Unitrode Power Supply Design Seminar Manual. SEM 800

[58] H Matsuo, N Aoike, F Kurokawa, A Hisako 1994 A new combined voltage-resonant inverter with high power factor and low distortion factor. *IEEE Power Electron. Spec. Conf., Rec.* 1:

[59] A Ferrari de Souza, I Barbi 1994 A new ZVS-PWM unity power factor rectifier with reduced conduction losses. *IEEE Power Electron. Spec. Conf., Rec.* 1:

[60] Gyu-Ha Choe, Min-Ho Park 1988 A new injection method for ac harmonic elimination by active power filter. *IEEE Trans. Ind. Electron.* 35: 141–147

[61] J Nastran, R Cajhen, M Seliger, P Jereb 1994 Active power filter for non-linear ac loads. *IEEE Trans. Power Electron.* 9: 92–96

[62] H Akagi 1994 Trends in active power line conditioners. *IEEE Trans. Power Electron.* 9: 263–268

[63] H-L Jou, J-C Wu, H-Y Chu 1994 New single-phase active power filter. *IEE Proc. Electron. Power Appl.* 141: 129–134

[64] D A Torrey, Adel M A M, Al-Zamel 1995 Single-phase active power filters for multiple non-linear loads. *IEEE Trans. Power Electron.* 10: 263–272

[65] C A Canesin, I Barbi 1996 Analysis and design of constant-frequency peak-current-controlled high- power-factor boost rectifier with slope compensation. *IEEE Appl. Power Electron. Conf., Rec.*, pp 807–813

[66] J Lazar, S Cuk 1996 Feedback loop analysis for ac/dc rectifiers operating in discontinuous conduction mode. *IEEE Appl. Power Electron. Conf., Rec.*, pp 797–806

[67] R B Ridley 1994 Average small-signal analysis of the boost power factor correction circuit. Volume V of the Virginia Power Electron. Center Publication Series

[68] F A Huliehel, F C Lee, B H Cho 1992 Small-signal modelling of the single -phase boost high power factor converter with constant frequency control. *IEEE Power Electron. Spec. Conf., Rec.*, pp 475–482

[69] P D Ziogas 1981 The delta modulation technique in static PWM converters. *IEEE Trans. Ind. Appl.* IA-17: 199–203

[70] M S Elmore, W A Peterson, S D Sherwood 1991 A power factor enhancement circuit. *IEEE Appl. Power Electron. Conf., Rec.*, pp 407–414

[71] R Oruganti, C Y Thean 1994 A novel PFC scheme for AC to DC converter with reduced losses. *IEEE Ind. Electron. Conf., Rec.*

[72] M B Harris, A W Kelly, J P Rhode, M E Baran 1994 Instrumentation for measurement of line impedance. *IEEE Appl. Power Electron. Conf., Rec.*, pp 887–893

[73] Valtkovic', D Borojevic', F C Lee 1996 Input filter design for power factor correction circuits. *IEEE Trans. Power Electron.* 11: 199–205

[74] F S Dos Reis, J Sebastia'n, J Uceda 1994 Determination of EMI emission in power factor preregulators by design. *IEEE Power Electron. Spec. Conf., Rec.* 2:

[75] M O Eissa, S B Leeb, G C Verghese, A M Stankovic 1996 Fast controller for a unity-power-factor PWM rectifier. *IEEE Trans. Power Electron.* 11: 1–6

[76] G Spiazzi, P Mattavelli, L Rossetto 1995 Power factor preregulators with improved dynamic performance. *IEEE Power Electron. Spec. Conf., Rec.*, pp 150–156

[77] Y Jiang 1994 Development of advanced power factor correction techniques. Ph D Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA

[78] D M Divan, G Venkataramanan, C Chen 1992 A unity power factor forward converter. *IEEE Ind. Appl. Soc., Annu. Mtg., Rec.*, pp 666–672

[79] I Takahashi, R Y Igarashi 1991 A switching power supply of 99% power factor by the dither rectifier. *IEEE Int. Telecommun. Energy. Conf., Rec.*, pp 714–719

[80] M H Kheraluwala, A R Schnore, R L Steigerwald 1992 Performance characterisation of a high power factor power supply with a single power stage. *IEEE Ind. Appl. Soc., Annu. Mtg, Rec.*, pp 659–665

[81] I Takahashi, R Y Igarashi 1992 High power factor switching regulator with no rush current. *IEEE Ind. Appl. Soc., Annu. Mtg., Rec.*, pp 673–680

[82] R Redl, L Balogh, N O Sokal 1994 A new family of single-stage isolated power-factor correctors with fast regulation of the output voltage. *IEEE Power Electron. Spec. Conf., Rec.* 2: 1137–1144

[83] M Madigan, R Erickson, E Ismail 1992 Integrated high quality rectifier-regulators. *IEEE Power Electron. Spec. Conf., Rec.*, pp 1043–1051

[84] R Redl, L Balogh 1992 RMS, DC, peak, and harmonic currents in high-frequency power-factor correctors with capacitive energy storage. IEEE catalog No. 0-7803-0485-3/92, pp 533–540

[85] R Redl, L Balogh 1995 Design considerations for single-stage isolated power-factor-corrected power supplies with fast regulation of the output voltage. *IEEE Appl. Power Electron. Conf., Rec.*, pp 454–458

[86] M J Willers, M G Egan, J M D Murphy, S Daly 1994 A BIFRED converter with a wide load range. *IEEE Ind. Electron. Conf., Rec.*, pp 226–231

[87] M M Javanovic, D M C Tsang, F C Lee 1994 Reduction of voltage stress in integrated high-quality rectifier-regulators by variable-frequency control. *IEEE Appl. Power Electron. Conf., Rec.*, pp 569–575

[88] Y S Lee, K W Siu 1996 Single-switch fast response switching regulators with unity power factor. *IEEE Appl. Power Electron. Conf., Rec.*, pp 791–796

[89] M Brkovic, S Cuk 1995 A novel single-stage AC-DC full-bridge converter with magnetic amplifiers for input current shaping. *IEEE Power Electron. Spec. Conf., Rec.*, pp 990–995

[90] Y Jiang, F C Lee, G Hua, W Tang 1993 A novel single phase power factor correction scheme. *IEEE Appl. Power Electron. Conf., Rec.*, pp 287–292

[91] Y Jiang, F C Lee 1994 Single-stage single-phase parallel power factor correction scheme. *IEEE Power Electron. Spec. Conf., Rec.* 2:

[92] J P Gegner, C-Y Hung, C Q Lee 1994 High power factor ac-to-dc converter using a reactive shunt regulator. *IEEE Power Electron. Spec. Conf., Rec.* 1: 349–355

[93] R Srinivasan, R Oruganti 1996 A single phase parallel power processing scheme with power factor control. *Int. J. Electron.* 80: 291–306 (Also in *IEEE Power Electron. Drive Syst. Conf., Rec.* 1: 1995)

[94] R Srinivasan, R Oruganti 1997 A single phase parallel power processing scheme with input-shunt power factor correction stage. *IEEE Power Electron. Drive Syst. Conf., Rec.*

[95] W-I Tsai, Y-Y Sun, W-S Shieh 1991 Modelling and control of single phase switching mode rectifiers with near-optimum dynamic regulation. *IEEE Ind. Electron. Conf., Rec.*, pp 501–506

[96] B R Needham, P H Eckerling, K Siri 1994 Simulation of large distributed DC power system using averaged modelling techniques and the SABER simulator. *IEEE Appl. Power Electron. Conf., Rec.*, pp 801–807

[97] D Kimhi, S Ben-Yaakov 1991 SPICE model for current mode PWM converters operating under continuous inductor current conditions. *IEEE Trans. Power Electron.* 6: 281–286

[98] K Mahabir, G Verghese, J Thottuvelil, A Heyman 1990 Linear averaged and sampled data models for large signal control of high power factor AC-DC converters. *IEEE Power Electron. Spec. Conf., Rec.*, pp 372–381

[99] S R Sanders, J M Noworolski, X Z Liu, G C Verghese 1991 Generalised averaging method for power conversion circuits. *IEEE Trans. Power Electron.* 6: 251–259

[100] J Thottuvelil, D Chin, G C Verghese 1991 Hierarchical approaches to modelling high-power-factor AC-DC converters. *IEEE Trans. Power Electron.* 6: 179–187

[101] P T Krein, J Bentsman, R M Bass, B L Lesieutre 1990 On the use of averaging for the analysis of power electronic systems. *IEEE Trans. Power Electron.* 5: 182–190

[102] L Dixon 1989 High power factor preregulators for off-line switching power supplies. Unitrode Power Supply Design Seminar Manual, SEM 600

[103] R Watson, G C Hua, F C Lee 1994 Characterisation of an active clamp flyback topology for dc-dc conversion and power factor correction alications. *IEEE Appl. Power Electron. Conf., Rec.*, pp 412–418

[104] R Watson, F C Lee, G C Hua 1994 Utilisation of an active-clamp circuit to achieve soft switching in flyback converters. *IEEE Power Electron. Spec. Conf. Rec.*

[105] G C Hua, C S Leu, Y M Jiang, F C Lee 1992 Novel zero-voltage-transition PWM converters. *IEEE Power Electron. Spec. Conf., Rec.*, pp 55–61

[106] R Srinivasan 1998 *Single phase power factor correction – investigation and development of advanced techniques*. Ph D thesis, National University of Singapore, Singapore

Flexible AC transmission systems: A status review

K R PADIYAR and A M KULKARNI

Department of Electrical Engineering, Indian Institute of Science,
Bangalore, India 560 012
e-mail: krpyar@ee.iisc.ernet.in

Abstract. With the availability of high power semiconductor switches with turn-off capability, voltage source converter based controllers for power transmission system applications have become a reality. Prototypes of some second generation Flexible AC Transmission System (FACTS) controllers like TCSC and STATCON have been installed. This paper presents a review of the progress in FACTS. A generalized description of FACTS controllers is also presented.

Keywords. Flexible AC transmission systems; static condenser; unified power flow controller; thyristor controlled series compensation.

1. Introduction

Expansion in power transmission networks has taken place not only due to the increase in generation and loads but also due to the extensive interconnection among different power utilities. The major factor for system interconnection is the economy possible in the reduced generation reserves for achieving the same level of reliability of supply.

Except for the limited number of HVDC links in a system, the vast majority of transmission lines are AC. The power flows in AC lines are uncontrolled and are determined by Kirchhoff's laws. This is in contrast to HVDC links where the power flow has to be regulated by converter controls. The lack of control in AC networks can be considered an advantage from the point of view of avoiding additional equipment. However AC lines have the following disadvantages.

(1) Power flow in parallel paths is determined according to their reactance. As a first approximation, the power flow in AC networks can be compared to the power flow in DC resistive networks where the resistance is analogous to the reactance. The operation of KVL (Kirchhoff's voltage law) implies that the network is often not optimally utilized.

(2) Power flow in AC lines (except short lines of lengths below 150 km) is limited by stability considerations. This implies that the lines may operate normally only at power levels much below their thermal limits.

781

(3) The operation of KVL and lack of control in AC lines implies that the normal power flow in a line is kept much below the peak value which itself is limited by stability (as mentioned earlier). This margin (or reserve) is required to maintain system security under contingency conditions.

(4) The AC transmission network requires dynamic reactive power control to maintain satisfactory voltage profile under varying load conditions and transient disturbances.

(5) AC lines, while providing synchronizing torque for oscillating generator rotors may contribute negative damping torque which results in undamped power oscillations, (particularly with fast acting static exciters and high gain automatic voltage regulators).

(6) The increase in load levels is accompanied by higher reactive power consumption in the line reactances. In case of mismatch in the reactive power balance in the system, this can result in voltage instability and collapse.

The reactive power compensation of AC lines using fixed series or shunt capacitors can solve some of the problems associated with AC networks. However the slow nature of control using mechanical switches (circuit breakers) and limits on the frequency of switching imply that faster dynamic controls are required to overcome the problems of AC transmission networks. Recent developments involving deregulation and restructuring of the power industry are aimed at isolating the supply of electrical energy (a product) from the service, involving transmission from generating stations to load centres. This approach is feasible only if the operation of AC transmission lines is made flexible by introducing fast-acting high-power solid state controllers using thyristor or GTO valves (switches). The advent of high voltage and high power thyristor valves and digital controllers in HVDC transmission has demonstrated the viability of deploying such controllers for power transmission. Thyristor controllers were also utilized in the late seventies to control current in reactors and switch capacitors and this led to the development of Static Var Compensators (SVC).

Flexible AC Transmission System (FACTS) is a concept proposed by Hingorani (1988, 1991, 1993) that involves the application of high power electronic controllers in AC transmission networks which enable fast and reliable control of power flows and voltages. FACTS do not indicate a particular controller but a host of controllers which the system planner can choose, based on cost benefit analysis. The objectives are as below.

(1) Regulation of power flows in prescribed transmission routes.

(2) Secure loading of lines nearer their thermal limits.

(3) Prevention of cascading outages by contributing to emergency control.

(4) Damping of oscillations which can threaten security or limit the usable line capacity.

Table 1 shows that the controllers can be broadly classified into two classes:

(a) shunt-connected controllers providing voltage control;

(b) series-connected controllers providing power flow control.

FACTS controllers.

	Type	Main function	Controller used	Comments
var compensator)	Shunt	Voltage control	Thyristor	Variable impedance device
ristor controlled ensation)	Series	Power flow control	Thyristor	Variable impedance device
yristor controlled regulator)	Series and shunt	Power flow control	Thyristor	Phase control using series (quadrature) voltage injection
(static condenser)	Shunt	Voltage control	GTO *	Variable voltage source
c synchronous ensator)	Series	Power flow control	GTO * source	Variable voltage
fied power flow	Shunt and series	Voltage and power flow control	GTO *	Variable voltage source

tead of the GTO, other power semiconductor devices with turn-off capability such as IGBT or S controlled thyristors) can also be used.

R provides power flow control mainly by injecting a quadrature voltage in se-
complex power generated by the series voltage source is supplied by the shunt-
d transformer. The simplified expression for power flow in a lossless transmission
en by

$$P = \frac{V_S V_R \sin(\delta_{SR} + \phi)}{X}, \tag{1}$$

and V_R are sending and receiving end bus voltages, X is the series reactance of
δ_{SR} is the difference in the bus angles, ϕ is the phase angle shift introduced by a
gle regulator (phase shifting transformer).
vious from (1) that the control of voltage, series reactance and phase angle (ϕ)
e power flow. While the control over the first two variables can be used to increase
r limit, the control over ϕ can be used to regulate power flow in loops.
tor controllers have also been employed or proposed for the following.

ynchronous Resonance (NGH) damping (Hingorani 1981).

ristor Controlled Dynamic Brake (Rao & Nagsarkar 1984; Raschio *et al* 1995).

Current Limiting (Salama *et al* 1993; Sugimoto *et al* 1996).

voltage Protection (Sarkozi *et al* 1994).

S controllers have also been proposed in distribution systems for control of power
Larsen *et al* 1992; Hingorani 1995; Akagi 1996; Sabin & Sundaram 1996). The
s are to limit voltage fluctuations and reduce the impact of momentary interrup-
would affect sensitive loads. In addition, distribution type FACTS devices can be

In this paper, the state of the art in the development of FACTS is presented. The attention is focussed on the new FACTS controllers based on voltage source converters (STATCON and UPFC).

2.   Description of FACTS controllers

In this section we briefly describe the basic principle of operation of some of the FACTS controllers.

2.1.   *Static Var compensator*

SVC is an important FACTS device already widely in operation. Ratings range from 60 MVAR to 600MVAR. SVC can be considered as a "first generation" FACTS controller and uses thyristor controllers. It is a shunt reactive compensation controller (Gyugyi 1979, 1988; Miller 1982) consisting of a combination of fixed capacitor or thyristor-switched capacitor in conjunction with thyristor-controlled reactor (FC-TCR or TSC-TCR) (figure 1). The SVC is a variable susceptance controller; the effective susceptance is varied by changing the conduction time of the thyristors of the TCR and/or switching in/out the shunt capacitor using a TSC. Early applications of the SVC were for load compensation of fast-changing loads like arc furnaces and steel mills for dynamic power factor improvement and load balancing in the three phases. Transmission line compensators are used not only for reactive power compensation (maintaining voltages at key points of the network within limits) but also for improving system stability. The control strategy usually employed is to use the SVC as a voltage regulator. In addition supplementary controls are used for damping power oscillations. The steady state control characteristics of a SVC is shown in figure 2. In the controllable range, a small slope is given to the characteristic to prevent frequent hitting of limits and also to facilitate parallel operation.
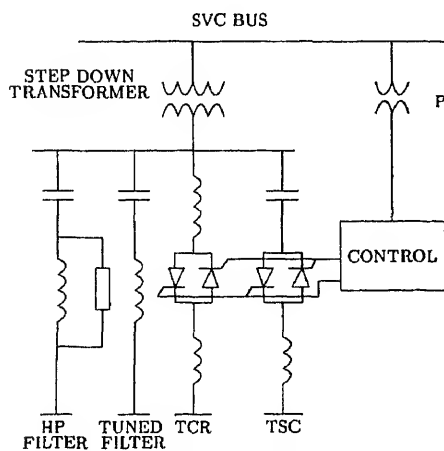
Figure 1.   TCR-TSC type SVC.
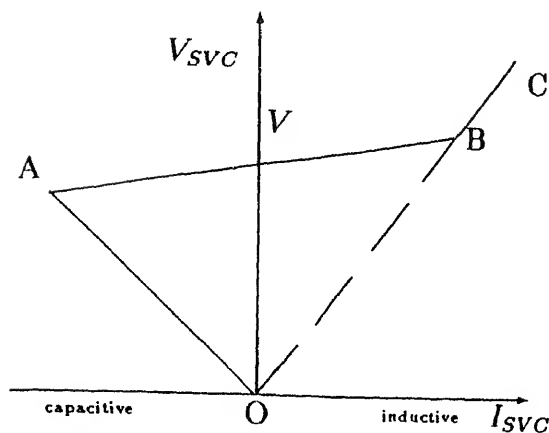
Figure 2. Controller characteristics of SVC.

2.2 *Thyristor-controlled series compensator*

The TCSC is a second generation FACTS controller which controls the effective line reactance by connecting a variable reactance in series with the line. The variable reactance is obtained using a FC-TCR combination with mechanically switched capacitor sections in series. Figure 3 shows a TCSC with two modules connected in series. Use of series compensation is usually a preferable alternative for increasing power flow capability of lines as compared to shunt compensators as the ratings required for series compensators are significantly smaller. Use of TCSC can reduce the adverse torsional interactions with generator-turbine shafts ("Subsynchronous Resonance") which is a major concern in the application of series capacitors.

The first demonstration project of TCSC was commissioned in 1991 at a 345 kV Kanawha River Substation in West Virginia, USA under American Electric Power Company. This was a test installation of thyristor switches in one phase for rapid switching of series capacitor segments and was supplied by Asea Brown Boveri, Sweden.

In October 1992, the first three-phase TCSC was installed at 230 kV Kayenta Substation in Arizona under the Western Area Power Administration (WAPA) (Christl 1991). Here a 15Ω capacitor bank is connected in parallel with a TCR and permits smooth and rapid control of (capacitive) reactance between 15 and 60 Ω through phase control of TCR (α varying between 145 to 180).
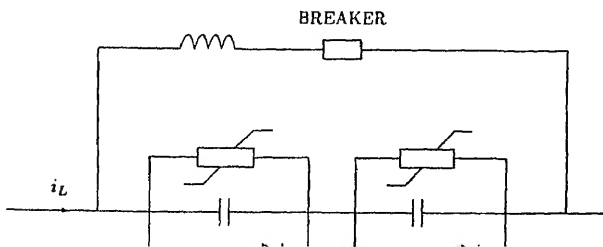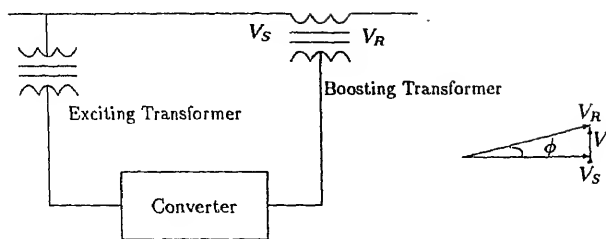
Figure 4.   Static phase shifter schematic.

A large prototype three-phase TCSC was installed in 1993 at 500 kV Slatt Substation in Oregon under the Bonneville Power Administration (BPA) (Urbanek *et al* 1993; Piwko *et al* 1994). The project was sponsored by the Electric Power Research Institute (EPRI) and the equipment was developed by General Electric in U.S.A. Here, six modules of TCSC are connected in series and controlled to provide a variation in impedance from $+1.4\Omega$ to -16Ω.

2.3   *Thyristor-controlled phase angle regulator*

This is also known as Static Phase Shifter (SPS) and phase shift with respect to the bus voltage is achieved by adding or subtracting a variable voltage component in quadrature with the bus voltage. This variable quadrature voltage component in one phase is obtained by thyristor switches and by an exciting transformer which is connected between the other two phases. The quadrature voltage is injected in series with the transmission line by a boosting transformer. The basic arrangement is shown in figure 4.

Many configurations of the SPS have been proposed (some employing GTOs). Also, with some configurations continuously variable phase angle of injected voltage (not necessarily in quadrature) is achievable. For a description of available topologies for SPS and a comparison between them see the review by Iravani & Maratukulam (1994).

2.4   *STATCON*

STATCON is a shunt reactive compensation device; however, unlike the SVC it is based on voltage source converter (VSC) using GTOs (figure 5). The principle of operation is similar to that of a synchronous condenser. The VSC is connected to the system through a small reactance which is the leakage reactance of the coupling transformer. The VSC produces a set of three phase voltages which are in phase with the corresponding bus voltages. The reactive power is varied by varying the magnitude of the converter output voltages. A small phase difference exists in the steady state (depending on reactive power output) so that real power can be drawn from the lines to compensate for the losses. The current on the DC side is mainly a ripple of magnitude much smaller than the AC line currents. As no real energy exchange (except to compensate for losses) takes place in steady state, the DC voltage can be maintained by a capacitor.

The STATCON can also exchange active power with the system if it has an energy source at its DC terminals. In this situation, the phase of the inverter output voltages with respect to corresponding bus voltages is varied to obtain control over active power.
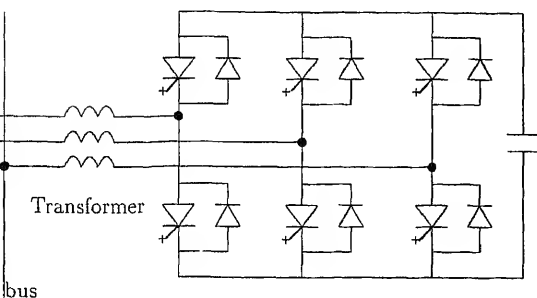
Figure 5. Six-pulse STATCON.

The major advantages of the STATCON over the SVC are (Edwards *et al* 1988; Larsen *al* 1992):

) The STATCON can supply required reactive current even at low values of bus voltage, whereas the reactive current capability of SVC at its susceptance limit decreases linearly with decrease in bus voltage (compare figures 2 and 6).

) With proper choice of device ratings and thermal design, STATCON can have a short time overload capability. This is not possible in an SVC because there is an inherent susceptance limit.

) Significant size reduction can be achieved because of reduced number of passive components and their smaller size.

) STATCON can allow for real power modulation if it has an energy source or large energy storage at its DC terminals.

An early application of a ±20 MVAR STATCON (Sumi *et al* 1981) used forced commutated thyristors. Subsequently there have been GTO based installations which include a 1 MVAR (Edwards *et al* 1988) and a larger ±80 MVAR (Mori *et al* 1993) STATCON. A rge ±100 MVAR prototype installation has been reported recently (Schauder *et al* 1995). me of the salient features of this STATCON are as follows.

) 48-pulse inverter (8 inverters connected in series on the AC side with appropriate phase shifts and in parallel on the DC side).

) DC capacitor voltage: 6.6 kV.
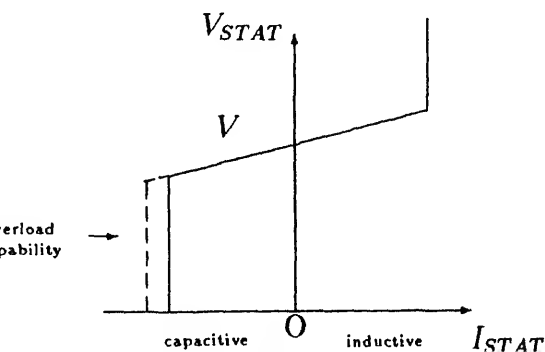
Figure 6. Control characteristics of STATCON.

(3) Inverter valves employ 5 GTO modules in series.

(4) GTO rating: 4500 V, 4000 A (peak turn-off).

A bidirectional 18-pulse voltage source converter utilizing GTOs and rated at 10 MVA was reported (Walker 1990). The converter is rated to operate in all four quadrants and connects a storage battery to a utility grid. It can be used for static Var control, as a power system stabilizer or as a real power peaking station.

In addition to multipulse topology, multilevel converter topology has also been proposed for the STATCON (Menzies & Zhuang 1995; Ekanayaki & Jenkins 1996). However the optimum circuit configuration is difficult to choose because relative significance of design parameters will vary from application to application.

2.5 *Unified power flow controller (UPFC)*

The UPFC (Gyugyi 1992, 1994, 1995) is the most versatile of FACTS controllers. The main function of the UPFC is to control real and reactive power flow through the line by injection of a (variable) voltage in series with the transmission line. A schematic of the UPFC is shown in figure 7. The UPFC consists of 2 GTO-based inverter branches. The series branch consists of a voltage source inverter which injects a voltage in series through a transformer. Since the series branch of the UPFC can inject a voltage with variable magnitude and phase angle it can exchange real power with the transmission line. However the UPFC as a whole cannot supply or absorb real power in steady state (except for the power drawn to compensate for the losses) unless it has a power source at its DC terminals. Thus the shunt branch is required to compensate (from the system) for any real power drawn/supplied by the series branch and the losses. In addition the shunt branch can exchange reactive power *independently* with the system. Thus the UPFC offers three controllable parameters. If only the series branch of the UPFC is used, there is only one controllable parameter which is the magnitude of the injected voltage (the phase of the injected voltage is constrained to be in quadrature with the line current so that the device cannot supply or absorb real power). This device is called the Static Synchronous Series Compensator (SSSC).

The world's first demonstration of a UPFC is under installation by American Electric Power Co., at the utility's Inez Station in eastern Kentucky, USA. It is to be installed in two phases; the ±160 MVAr shunt inverter has been installed first for voltage support in 1997. Subsequently, the series inverter of the same rating will be installed. The converter will use GTOs and a three-level topology.
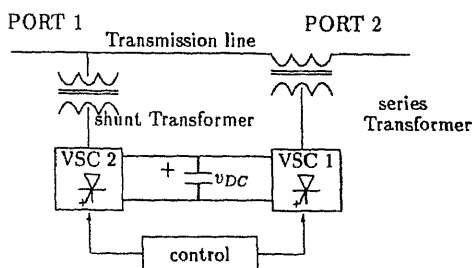
Figure 7.   UPFC.
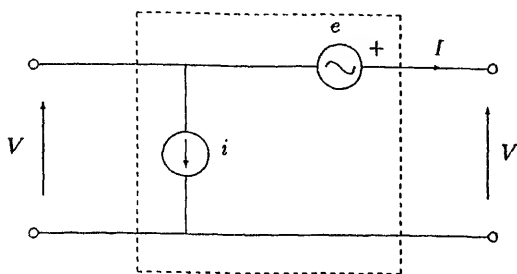
Figure 8. Circuit representation.

2.6 Generalized description

Based on the treatment given by Gyugyi (1992), the influence of a FACTS controller in a transmission line can be viewed as injection of a shunt current source and a series voltage source (see figure 8). This is the most general description of a FACTS controller, such as a UPFC.

Ignoring active power losses in the controller, the following constraint equation applies

$$Re[Vi^*] = Re[eI^*]. \tag{2}$$

It is assumed that both e and i are sinusoidal and can be expressed as phasors in steady state. Equation (2) shows that there are three independant variables (for example, magnitude and phase angle of e and the value of reactive component of current in i) that can be manipulated to control power flows (active and reactive) in a line.

It is to be noted that the general case, a FACTS controller may contain an energy source and thus even the constraint equation (2) does not apply. However this is unlikely except in the case of STATCON which may use a battery energy storage system for providing electrical energy during interruptions caused by system faults.

Except for UPFC, most of the FACTS controllers use only single control variable as listed in table 2.

Table 2. Mathematical description.

Controller	Constraint equations	Control variable
SVC	$e = 0, i = jB_{SVC}V$	B_{SVC}
TCSC	$i = 0, e = jX_{TCSC}I$	X_{TCSC}
TCPAR	$Vi = eI$	ϕ
	$V = Ve$	
STATCON without energy source	$e = 0, Re[Vi^*] = 0$	i (reactive current)
STATCON with energy source	$e = 0$	i_a (active current) i

3. Control of STATCON and UPFC

3.1 *STATCON control*

The primary controller of the STATCON is the reactive current (component of current in *quadrature* with the bus voltage) controller. A closed-loop control of reactive current is necessary because the STATCON current is dependent not only on the firing angle of the GTOs but also on the parameters of the rest of the system as well.

There are two controller structures which have been proposed for reactive current control of STATCON (Schrouder & Mehta 1993).

(1) *Type I*: The controller uses *both* magnitude and phase angle control of the voltage controllers to control the real and reactive current (figure 9). The real current is used to maintain capacitor at a constant voltage. The reactive current can be used as an inner loop of a bus voltage regulator. The real current is used to maintain capacitor voltage constant.

(2) *Type II*: Reactive current control can also be achieved by phase-angle control alone. The capacitor voltage is not controlled but depends upon the phase difference between the converter output voltages and the bus voltage (the phase difference is usually very small). Thus the converter voltage magnitude (which is dependent on the capacitor magnitude alone in this case) varies with phase angle. As a result, control over reactive current is possible by phase control. Control of bus voltage is achieved as in type I type controller (figure 10).

PI control of STATCON (type II) poses problems due to nonlinearity in the state equations. Nonlinear state feedback control is used to overcome this problem by Schrauder & Mehta (1993). Fuzzy logic control (which can be thought of as nonlinear heuristic control) has also been used in (Padiyar & Kulkarni 1997) and shows superior performance under varying system conditions. Control of STATCON for reducing overcurrents during unbalanced faults using PWM has been discussed by Jiang & Ekstrom (1995). In the design of control systems, the possibility of adverse interactions with the network and/or other FACTS controllers in the vicinity has to be kept in mind (Clark *et al* 1995; Woodford 1996). The problem of network-controller interactions has been reported in the voltage control of STATCON (Padiyar & Kulkarni 1997) which is solved by filtering the voltage signal.
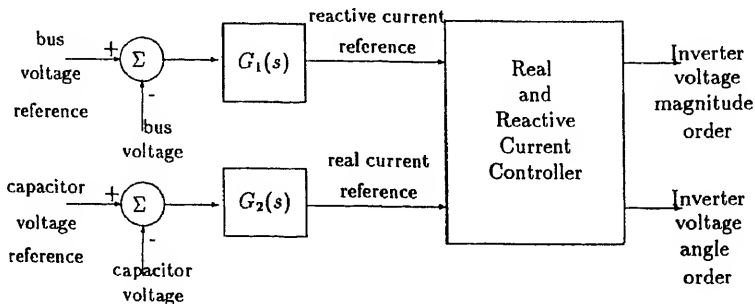
Figure 9. Type-I type controller.
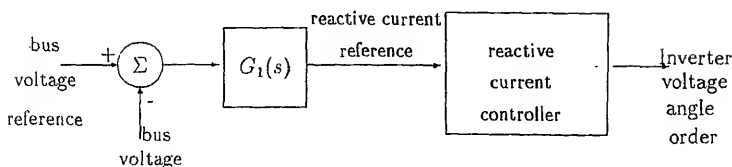
Figure 10. Type II type controller.

2 UPFC control

The UPFC allows three controllable parameters, viz., magnitude and phase of series injected voltage, and shunt reactive current. A control strategy for UPFC has been proposed (Padiyar & Kulkarni 1996) in which active power flow through the line is controlled, while regulating magnitudes of the voltages at its two ports.

The controller for this purpose uses only local measurements. The series voltage of the UPFC is split into two parts, one in phase with the line current ("real voltage") and the other in quadrature ("reactive voltage"). Series reactive voltage is used to control active power flow in the line, while real voltage and shunt reactive current injection are used to regulate voltage at the two ports. A block diagram representation of the series voltage controller is shown in figure 11. The power controller sets the reactive voltage reference. An auxiliary current angle stabilizing signal is used to damp a critical *network* mode associated with series reactance of line (steady state contribution of this auxiliary signal is zero because of the washout block $[sT_w/(1 + sT_w)]$). A damping controller for the slow *electromechanical* modes of the system associated with generator rotor swings modulates the power reference, and uses the generator slip signal. A synchronising component (K/s) is present in the controller in addition to the damping component (D) to reduce angle deviations. The real voltage to be injected can be calculated so as to maintain port 2 voltage magnitude of the UPFC constant. The voltage and power references can be set
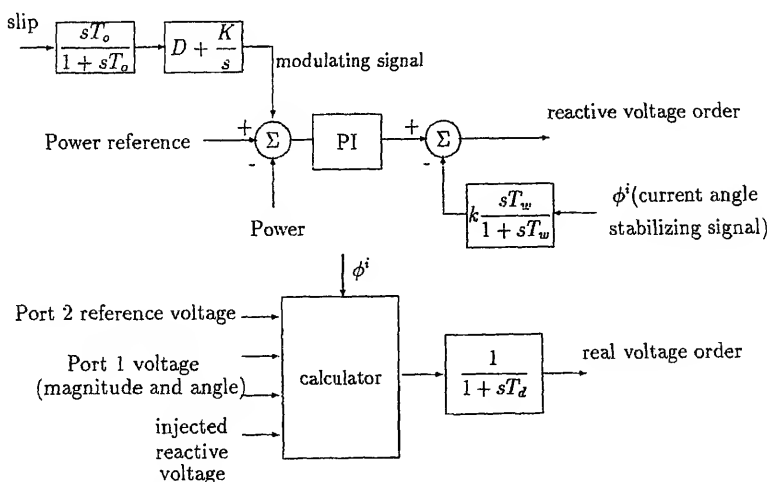
Figure 11. UPFC series voltage controller structure.

according to steady state real and reactive power flow requirements. The shunt current controller is similar to the one shown in figure 9.

4. Stability improvement using FACTS controllers

The fast response of FACTS controllers can be used effectively to improve dynamic performance of the system. The dynamic problems associated with power systems can be characterised as (Padiyar 1996):

(1) *Angle instability*: which includes small signal oscillatory instability ("dynamic instability") and large disturbance instability leading to loss of synchronism amongst generators ("transient instability"). These problems are associated with generator rotor swings.

(2) *Voltage instability*: which implies an uncontrolled collapse in voltage at load buses precipitated due to some disturbance and is primarily caused by dynamics of the load.

(3) *Subsynchronous resonance (SSR)*: This problem is caused by the adverse interaction of the electrical system with the generator-turbine mechanical system causing oscillatory instability or transient torques. This problem is usually present in radial series compensated lines connected to turbine generators.

While the application of FACTS controllers for steady state active and reactive power flow control is relatively straightforward, their use for stability improvement poses the following questions:

(1) Where should the controllers be located in the system for maximum effectiveness?

(2) What should be the control laws?

Usually FACTS controllers are used for steady state functions which is the main control objective, but supplementary controls are included for stability improvement. However, controllers such as Thyristor Controlled Brake (Rao & Nagsarkar 1984: Raschio *et al* 1995) and NGH-SSR damper (Hedin *et al* 1981; Hingorani 1981; Hingorani *et al* 1987; Benko *et al* 1987) are used specifically for transient stability improvement and SSR mitigation respectively. It should be kept in mind that the control strategy employed will vary from application to application depending on the relative severity of the various dynamic and steady state problems.

For SVCs, while voltage control is the main function, Supplementary Modulation Controller (SMC) can be introduced for damping small signal oscillations. This controller modulates either the voltage reference of the voltage controller or the susceptance. A typical configuration of SMC is shown in figure 12. A washout block is used to drive SMC output to zero in steady state. Similar control structures can be used for TCSC.

FILTER	GAIN	WASHOUT	PHASE LEAD

$u \longrightarrow \dfrac{1}{1+sT_F} \longrightarrow K_{SMC} \longrightarrow \dfrac{sT_W}{1+sT_W} \longrightarrow \dfrac{(1+T_1)(1+T_3)}{(1+T_2)(1+T_4)} \longrightarrow V$
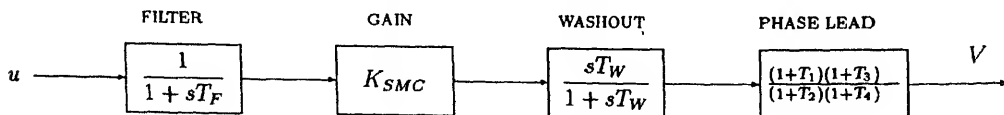
Figure 12. Supplementary modulation controller.

Methods proposed in literature for deciding location of FACTS controllers and control strategies (choice of control signal (u in the SMC) and the corresponding control law) are based mainly on eigenvalue sensitivity/residue (of transfer function) approaches (Padiyar *et al* 1986; Martins & Lima 1990; Okamato *et al* 1995). Computation of several indices has been proposed (Larsen *et al* 1995) to select the appropriate control signal and control law. Control strategies have also been proposed based on energy function approach (Gronquist *et al* 1995).

Study of large disturbance dynamics involves analysis of a large set of nonlinear differential equations. Design of control strategies and deciding on the best location from the point of view of transient stability improvement is a difficult problem and has to be validated using simulation studies. Large disturbance stability improvement with FACTS controllers has been shown using simulation studies (see for example Mihaltic *et al* 1996). Discrete control strategies have been proposed by Kosterev & Kolodziej (1995) and Padiyar & Uma Rao (1997).

The problem of torsional interactions (SSR) has been a major concern, primarily because some FACTS controllers like SVC have the potential to cause adverse interations, even in the absence of series compensation (Rustomkolai *et al* 1990). Therefore thorough studies need to be done in order that FACTS controllers themselves do not aggravate this problem. Studies on application of TCSC and field tests indicate that SSR problem can be reduced with it (see Piwko *et al* 1996). It is possible (FACTS controllers can have time of response of the order of 1-3 cycles) that supplementary controls be introduced to damp SSR directly. Use of auxiliary signals to damp SSR has been proposed for SVCs (see Ramey *et al* 1981, Vascjnczuk 1981, Putman & Ramey 1982, Hammad & El-Sadek 1984, Abi Samra *et al* 1985, Padiyar & Varma 1990). Further work needs to be carried out with new controllers such as STATCON, UPFC and SSSC.

Conclusions

The application of FACTS controllers throws up new challenges for power engineers, not only in hardware implementation, but also in design of robust control systems, planning and analysis.

There has been considerable progress in the application of FACTS controllers. Notably, prototypes of TCSC and STATCON of sufficiently high ratings are now in operation. Implementation of several other controllers such as UPFC are being planned. Field experiences with these installations are eagerly awaited by power engineers.

The new generation of FACTS controllers are voltage source converter-based and have several advantages over thyristor-controlled variable impedance type controllers. However, these controllers require switching devices with turn-off capability. At present, only GTO devices are available at the voltage and current ratings required. These devices have certain drawbacks and it is anticipated that in future, better and more efficient devices like IGBT and MCT will be available in the required ratings. Spurred by demand, the cost of these devices is also expected to come down.

There is every reason to believe that in a decade or so, FACTS controllers will revolutionize electrical power transmission systems making them more reliable, optimally utilized and better controlled.

References

Akagi H 1996 New trends in active filters for power conditioning. *IEEE Trans. Ind. Appl.* 32: 1312–1322

Abi Samra N C, Smith R F, McDermott T E, Chidester M B 1985 Analysis of thyristor controlled shunt SSR countermeasures. *IEEE Trans. Power Apparatus Syst.* 104: 584–597

Benko I S, Bhargava B, Rothenbuler W N 1987 Prototype NGH subsynchronous resonance damping scheme. Part II - Switching and short circuit tests. *IEEE Trans. Power Syst.* 12: 1040–1049

Clark K, Fardanesh B, Adapa R 1995 Thyristor controlled series compensation application study-Control interaction considerations. *IEEE Trans. Power Delivery* 10: 1031–1037

Christl N, Hedin R, Johnson R, Krause P, Montoya A 1991 Power system studies and modelling for Kayenta 230 kV substation. Advanced series compensation. In *Fifth Int. Conf. on AC and DC Transmission* London, *IEE Conf. Publ. No.* 345: 33–37

Ekanayake J B, Jenkins N 1996 A three level advanced static var compensator. *IEEE Trans. Power Delivery* 11: 540–545

Edwards C W, Nannery P R, Mattern K E, Stacey E J, Gubernick J 1988 Advanced static var generator employing GTO thyristors. *IEEE Trans. Power Delivery* 3: 1622–1627

Gronquist J F, Sethares W A, Alvarado F L, Lasseter R H 1995 Power oscillation damping strategies for FACTS devices using locally measurable quantities. *IEEE Trans. Power Syst.* 10: 1598–1605

Gyugyi L 1979 Reactive power generation and control using thyristor circuits. *IEEE Trans. Ind. Appl.* 15: 521–531

Gyugyi L 1988 Power electronics in electrical utilities: Static var compensators. *Proc. IEEE* 76: 482–594

Gyugyi L 1992 Unified power flow control concept for flexible AC transmission systems. *Inst. Elec. Eng. Proc.* C139: 323–331

Gyugyi L 1994 Dynamic compensation of AC transmission lines by solid state synchronous voltage sources. *IEEE Trans. Power Delivery* 9: 904–911

Gyugyi L 1995 The unified power flow controller: A new approach to power transmission control. *IEEE Trans. Power Delivery* 10: 1085–1099

Hammad A E, El-Sadek M 1984 Application of a thyristor controlled VAR compensator for damping subsynchronous oscillations in power systems. *IEEE Trans. Power Apparatus Syst.* 103: 198–212

Hingorani N G 1981 A new scheme for subsynchronous resonance damping of torsional oscillations and transient torque – Part I. *IEEE Trans. Power Apparatus Syst.* 100: 1852–1855

Hingorani N G 1988 Power electronics in electrical utilities: Role of power electronics in future power systems. *Proc. IEEE* 76: 481–482

Hingorani N G 1991 FACTS–Flexible AC transmission systems. In *Fifth Int. Conf. on AC and DC Transmission*, London *IEE Conf. Publ.* No. 345: 1–7

Hingorani N G 1993 Flexible AC transmission systems *IEEE Spectrum* 30: 40–45

Hingorani N G 1995 Custom power *IEEE Spectrum* 32: 41–48

Hingorani N G, Bhargava B, Garrigue G F, Rodriguez G D 1987 Prototype NGH subsynchronous resonance damping scheme–Part I - Field installation and operating experience. *IEEE Trans. Power Syst.* 12: 1034–1039

Hedin R A, Stump K B, Hingorani N G 1981 A new scheme for subsynchronous resonance damping of torsional oscillations and transient torque – Part II. *IEEE Trans. Power Apparatus Syst.* 100: 1856–1863

Iravani M R, Maratukulam D 1994 Review of semiconductor controlled phase shifters for power system applications. *IEEE Trans. Power Syst.* 9: 1833–1839

Jiang Y, Ekstrom A 1995 Applying PWM to control overcurrent at unbalanced faults of force commutated voltage source converters used as static var compensators. In *Stockholm Power Technol. Conference*, Stockholm, Sweden, pp 18–22

Kosterev D N, Kolodziej W J 1995 Bang-bang series capacitor transient stability. *IEEE Trans. Power Syst.* 10: 915–923

Larsen E V, Leonard D J, Miller N W, Othmann H, Paserba J J, Naumann S J 1992 Application studies for a distribution STATCON on commonwealth Edison's power system. In *Proceedings of FACTS Conference*, Electric Power Res. Inst. (EPRI), Boston

Larsen E V, Miller N, Nilsson S, Lindgren S 1992 Benefits of GTO–based compensation systems for electric utility application. *IEEE Trans. Power Delivery* 7: 2056–2064

Larsen E V, Sanchez-Gasca J J, Chow J H 1995 Concept for design of FACTS controllers to damp power swings. *IEEE Trans. Power Syst.* 10: 948–956

Miller T J E 1982 *Reactive power control in electric systems* (New York: John Wiley)

Martins N, Lima L T G 1990 Determination of suitable location for power system stabilizers and static var compensators for damping electromechanical oscillations in large scale systems. *IEEE Trans. Power Syst.* 5: 1455–1469

Mori S, Matsuno K, Hasegawa T, Ohnishe S, Takeda M, Seto S, Murakami S, Ishiguro F 1993 Development of large static var generator using self-commutated inverter for improving system stability. *IEEE Trans. Power Syst.* 8: 371–377

Menzies R W, Zhuang Y 1995 Advanced static compensation using multilevel GTO inverter. *IEEE Trans. Power Delivery* 10: 732–737

Mihalic R, Zunko P, Povh D 1996 Improvement of transient stability using unified power flow controller. *IEEE Trans. Power Delivery* 11: 485–491

Okamato H, Kurita A, Sekine Y 1995 Method for identification of effective locations of variable impedance apparatus on enhancement of steady state stability in large scale power systems *IEEE Trans. Power Syst.* 10: 1401–1407

Padiyar K R 1996 *Power system dynamics – Stability and control* (Bangalore: Interline)

Padiyar K R, Kulkarni A M 1997 Design of reactive current and voltage controller of static condenser. *Int. J. Elect. Power Energy Syst.* to appear

Padiyar K R, Kulkarni A M 1996 Development and evaluation of controls for unified power flow controller. In *Ninth National Power Systems Conference*, Indian Institute of Technology Kanpur, pp 253–257

Putman T H, Ramey D G 1982 Theory of modulated reactance solution for subsynchronous resonance. *IEEE Trans. Power Apparatus Syst.* 101: 1527–1535

Padiyar K R, Rajashekharam P, Radhakrishnan C, Pai M A 1986 Dynamic stabilization of power systems through reactive power modulation. *Elect. Mach. Power Syst.* 11: 281–293

Padiyar K R, Uma Rao K 1997 Discrete control of series compensation for stability improvement of power system. *Int. J. Elec. Power Energy Syst.* 19: 311–319

Padiyar K R, Varma R K 1990 Static var system auxiliary controllers for damping torsional oscillations. *Int. J. Elec. Power Energy Syst.* 12: 271–286

Piwko R J, Wegner C A, Furumasu B C, Damsky B L, Eden J D 1994 The Slatt thyristor controlled series capacitor project: Design, installation, commissioning and system testing. *Int. Conf. Large High Voltage Electric Systems (CIGRE)*, Paris, 14–104

Piwko R J, Wegner C A, Kinney S J, Eden J D 1996 Subsynchronous resonance performance tests of the slatt thyristor controlled series capacitor. *IEEE Trans. Power Delivery* 11: 1112–1119

Ramey D G, Kimmel D S, Dorney J W, Kroening F H 1981 Dynamic stabilizer verification tests at the San Juan station. *IEEE Trans. Power Apparatus Systems* 100: 5011–5019

Raschio P, Mittelstadt W A, Haner J F, Spee R, Enslin J H R 1995 Evaluation of dynamically controlled brake for western power system. In *CIGRE 1995 Symposium on Power Electronics in Electric Power Systems*, Tokyo, pp 22–24

Rao C S, Nagsarkar T K 1984 Half wave thyristor controlled dynamic brake to improve transient stability. *IEEE Trans. Power Apparatus Syst.* 103: 1077–1083

Rostomkolai N, Piwko R J, Larsen E V, Fischer D A, Mobarak M A, Poitras A E 1990 Subsynchronous torsional interaction with SVCs – Concepts and practical implications. *IEEE Trans. Power Syst.* 5: 1324–1332

Sabin D D, Sundaram A 1996 Quality enhances. *IEEE Spectrum* 33: 34–41

Salama M M A, Temraz H, Chikhani A Y, Bayoumi M A 1993 Fault current limiter thyristor controlled impedance. *IEEE Trans. Power Delivery* 8: 1518–1528

Sarkozi M, Gyugyi L, Bronfeld J D, Nilsson S, Damsky B 1994 Thyristor switched ZNO voltage limiter. *Proc. of Int. Conf. Large High Voltage Electric Systems (CIGRE)*, Paris, 14–302

Schauder C, Gernhardt M, Stacey E, Cease T W, Edris A, Lemak T, Gyugyi L 1995 Development of ±100 Mvar static condenser for voltage control of transmission systems. *IEEE Trans. Power Delivery* 10: 1486–1496

Schauder C, Mehta H 1993 Vector analysis and control of advanced static var compensator. *Inst. Elec. Eng. Proc.* C 140: 299–306

Sumi Y, Harumoto Y, Hasegawa T, Yano M, Ikeda K, Matsuura T 1981 New static var control using force commutated inverters. *IEEE Trans. Power Apparatus Syst.* 100: 4216–4224

Sugimoto S, Kida J, Arita H, Fukui C, Yamagiwa T 1996 Principle and characteristics of a fault current limiter with series compensation. *IEEE Trans. Power Delivery* 11: 842–847

Urbanek J, Piwko R J, Larsen E V, Damsky B L, Furumasu B C, Mittelstadt W, Eden J D 1993 Thyristor controlled series compensation prototype installation at the 500kV Slatt substation. *IEEE Trans. Power Delivery* 8: 4460–4469

Walker L 1990 A 10 MW GTO converter for battery peaking service. *IEEE Trans. Ind. Appl.* 26: 63–72

Wasynczuk O 1981 Damping subsynchronous resonance using reactive power control. *IEEE Trans. Power Apparatus Syst.* 100: 1096–1104

Woodford D A 1996 Electromagnetic design considerations for fast acting controllers. *IEEE Trans. Power Delivery* 11: 1515–1521

Advances in vector control of *ac* motor drives – A review

A K CHATTOPADHYAY

Department of Electrical Engineering, Bengal Engineering College (DU), Howrah 711 103, India
e-mail: ee@becs.ernet.in

Abstract. This paper attempts to present a comprehensive review of the advances made in vector control or field orientation as applied to high performance *ac* motor drives. Brief application survey, machine models in d-q representation, implementation issues with inverters and cycloconverters, parameter effects etc for both induction and synchronous motor vector control are dealt with and sample results from studies on them are presented. The latest advance on this control like direct torque control (DTC) has been briefly discussed. A substantial updated bibliography, though by no means complete, is included for those who are interested in keeping track of the present state-of-the-art and working further in this area.

Keywords. Vector control; field orientation; *ac* motor drives; high performance drives; induction motor; synchronous motor; direct torque control.

1. Introduction

Electric drives for motion control must have a fast torque response, four quadrant operation capability and controllability of torque and speed over a wide range of operating conditions. A separately excited *dc* motor, earlier used as the primary machine and later with simple power electronic controllers and current feedback, provides direct control of the magnitude of armature current and, in proportion, the torque, and has been the most popular choice for many industrial drives for such requirements in spite of its inherent drawback of the bulky, expensive and maintenance-prone commutator. On the other hand, *ac* motors, specially induction motors with their simple, less expensive, and more robust structures are more suitable for industrial environments though their control is quite complex. This is due to the fact that the rotor current in an induction motor which is responsible for the torque production owes its origin to the stator current which also contributes to the air-gap flux resulting in a coupling between the torque- and flux-producing mechanisms. In the *dc* machine, the field current in the stationary poles producing the magnetising flux and the armature current directly controlling the torque are independently accessible. Moreover, for a fully compensated *dc* motor, the spatial angle between the flux and the armature mmf is held at 90° with respect to each other, independent of the load, by the commutator and

the brushes whereas in an *ac* motor (both induction and synchronous), the spatial angle between the rotating stator and rotor fields varies with the load and gives rise to oscillatory dynamic response. Control methods for *ac* motors that emulate the *dc* motor control by orienting the stator current so as to attain independent and 'decoupled' control of flux and torque are known as 'field orientation' control and require control of both the magnitude and phase of *ac* quantities and thus are referred to as 'vector control methods'.

Early conceptual works in vector control were by Blaschke (1972) and Hasse (1969), which were translated into practical implementation later by Gabriel *et al* (1980), Leonhard (1985) and many others with the advances in microprocessors and microcomputers along with power electronics. Now, it has been established as a powerful technique in the field of *ac* motor drives and adopted worldwide. An exhaustive list of publications has been reported in this topic, which includes an IEEE Tutorial Course (Novotny & Lipo 1985) and two exclusive books (Vas 1990; Boldea & Nasar 1992). Work has continued unabated in this field and several issues like simplification of practical system with advanced microprocessors, design of current regulators/flux observers, reliability enhancement, performance improvement, parameter adaptation etc. are still attracting the researchers in this field. This paper attempts to make a summary review of the progress in vector control as applied to both induction and synchronous motor drives highlighting some typical results from the drives developed by the author and his research students at the Indian Institute of Technology, Kharagpur.

2. Vector control of induction motors

2.1 *Brief application survey*

The principle of vector control is used in current regulated PWM inverter (CRPWM), CSI, VSI, and cycloconverter-fed induction motor drives. The controlled current operation of the motor results in simpler implementation. The CRPWM inverter is common for high performance servo drives while CSI and cycloconverters are used for larger drives. High frequency PWM transistor inverters (10 kHz), developed around 1979, made it possible to use vector controllers in various kinds of industries including pinch roll drives of continuous casting plates, machine-tool drives and gear-less servo drives as reported by Kume & Iwakane (1987). The control method was applied to a large-scale paper mill (Tanaka *et al* 1983) with induction motors of 300–560 kW rating using CSI. Application of vector controlled induction motors for high performance servo drives has been brilliantly surveyed by Leonhard (1986). High horsepower vector controlled induction motor servo drive using adaptive rotor flux observer has been recently developed with improved steady state and dynamic response (Huang *et al* 1994). The recent trend is to eliminate the speed and position sensors in high performance vector controlled induction motor drives (Okuyama *et al* 1990; Onishi *et al* 1994; Tajima *et al* 1995). Very high power (MW) range cycloconverter-fed induction motors, with vector control for steel mill drive are mature drive systems in Japan (Sugi *et al* 1983; Saito *et al* 1987) and Germany (Timpe 1982; Hasse 1977). *Siemens* has recently announced optimised vector controlled SIMOVERT master drives for elevator applications (Scheirling & Schonherr 1995) having many important features.

2.2a *Dynamic model:* A dynamic model developed either with the concept of space phasors (Leonhard 1985; Murphy & Turnbull 1988) or *d-q* representations (Novotny & Lipo 1985; Bose 1986) may be utilised to develop the basic machine equations for implementation of vector control. We like to use the latter for convenience and familiarity. The *d-q* axes model of an induction motor with reference axis rotating at synchronous speed w_e is

$$
\begin{bmatrix} v_{ds}^e \\ v_{qs}^e \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} R_s + \sigma L_s p & -\sigma L_s \omega_e & \frac{L_m}{L_r'} p & -\frac{L_m}{L_r'} w_e \\ \sigma L_s \omega_e & R_s + \sigma L_s p & \frac{L_m}{L_r'} w_e & \frac{L_m}{L_r'} p \\ -L_m \frac{R_r}{L_r'} & 0 & \frac{R_r'}{L_r'} + p & -\omega_{sl} \\ 0 & -L_m \frac{R_r'}{L_r'} & \omega_{sl} & \frac{R_r'}{L_r'} + p \end{bmatrix} \begin{bmatrix} i_{ds}^e \\ i_{qs}^e \\ \psi_{dr}^{e'} \\ \psi_{qr}^{e'} \end{bmatrix}, \tag{1}
$$

where

$$
\sigma = 1 - \frac{L_m^2}{L_s L_r'}, \quad p = \frac{d}{dt}, \quad \omega_{sl} = (\omega_e - \omega_r).
$$

The electromagnetic torque developed by a 3-phase, *P*-pole, induction motor is

$$
T_e = \frac{3}{2} \frac{P}{2} \frac{L_m}{L_r'} (i_{qs}^e \psi_{dr}^{e'} - i_{ds}^e \psi_{qr}^{e'}), \tag{2}
$$

where

$$
\psi_{dr}^{e'} = L_m i_{ds}^e + L_r' i_{dr}^{e'}, \tag{3}
$$

$$
\psi_{qr}^{e'} = L_m i_{qs}^e + L_r' i_{qr}^{e'}. \tag{4}
$$

The field orientation implies that the stator current components obtained be oriented in phase (flux component) and in quadrature (torque component) to the flux vector which can be either stator flux (ψ_s), airgap or mutual or magnetising flux (ψ_m), or rotor flux (ψ_r) as shown in the equivalent circuit in figure 1 (Sathiakumar *et al* 1986). The orientation of the stator current with respect to the stator, rotor and airgap flux has been examined and the relative merits and developments of the schemes have been reported (Bayer & Blaschke 1977; Sathiakumar *et al* 1986; Ho & Sen 1988; Erdman & Hoft 1990). It has been shown that the rotor flux orientation alone provides natural decoupling, fast torque response and all round stability. The stator flux and airgap flux orientation, however, are attractive due to ease of flux computation and for the purpose of wide range of field weakening operation (Xu & Novotny 1992) but need decoupler network (De Doncker & Novotny 1988). A new strategy called the 'Universal field oriented controller' has been developed by De Doncker & Novotny (1988) which decouples flux and torque in an arbitrary flux reference frame.

Rewriting the rotor voltage equations in (1)

$$
p\psi_{dr}^{e'} + \frac{R_r'}{L_r'} \psi_{dr}^{e'} - \frac{L_m}{L_r'} R_r' i_{ds}^e - \omega_{sl} \psi_{qr}^{e'} = 0, \tag{5}
$$

$$
p\psi_{qr}^{e'} + \frac{R_r'}{L_r'} \psi_{qr}^{e'} - \frac{L_m}{L_r'} R_r' i_{qs}^e + \omega_{sl} \psi_{dr}^{e'} = 0. \tag{6}
$$

$$L_{ls} = L_s - L_m, \quad L'_{lr} = L'_r - L_m$$
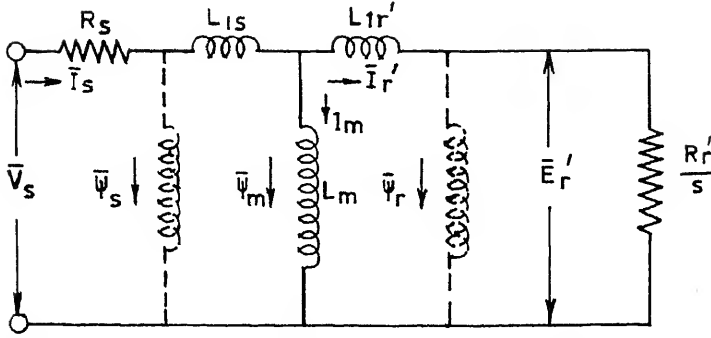
Figure 1. Conventional stator referred induction motor equivalent circuit showing different flux vectors.

For rotor flux orientation control, the rotor flux axes are locked with the synchronously rotating reference system such that the rotor flux is entirely in the d-axis,

$$\psi_r^{e'} = \psi_{dr}^{e}{}', \tag{7}$$

$$\psi_{qr}^{e'} = 0. \tag{8}$$

Substituting (7) & (8) in (5) & (6) yields

$$\omega_{sl} = \frac{L_m}{\psi_r^{'e}} \left(\frac{R'_r}{L'_r} \right) i_{qs}^e, \tag{9}$$

$$\frac{L'_r}{R'_r} p\psi_r^{'e} + \psi_r^{'e} = L_m i_{ds}^e. \tag{10}$$

For the range of operation below the base speed, the flux $\psi_r^{e'} = \psi_{dr}^{e'}$ is kept constant, when

$$p\psi_{dr}^{e'} = 0. \tag{11}$$

From (4) & (8),

$$i_{qs}^e = -\frac{L'_r}{L_m} i_{qr}^{'}, \tag{12}$$

which shows a direct equilibrium relation between the torque component current i_{qs}^e and the rotor current $i_{qr}^{'e}$. The torque equation is

$$T_e = \frac{3}{2} \frac{P}{2} \frac{L_m}{L'_r} i_{qs}^e \psi_r^{'e}, \tag{13}$$

which shows the desired property of providing a torque proportional to the torque command i_{qs}^e.

During flux changes in the transient, $p\psi_{dr} \neq 0$ and from (10)

$$i_{ds}^e = \frac{\psi_{dr}^{'e} - L_m i_{ds}^e}{L'_r}. \tag{14}$$

Combining (14), (4) & (8) to eliminate $i_{dr}^{e'}$, yields the equation relating i_{ds}^e and $\psi_{dr}^{e'}$ (flux command and the flux),

$$(R_r' + L_r'p)\psi_{dr}^{e'} = R_r' L_m i_{ds}^e, \tag{15}$$

which in the steady state is

$$\psi_{dr}^{'e} = L_m i_{ds}^e. \tag{16}$$

The close parallel to the dc machine is now clearly visible. With the flux command held constant, a change in i_{qs}^e is followed instantly by corresponding change in $i_{qr}^{'e}$. While with a change in flux command, a transient rotor current is induced which subsequently decays with the rotor open circuit time constant L_r'/R_r' as shown in (15).

2.2b *Steady state model:* A convenient steady-state equivalent circuit model of the field oriented induction motor as shown in figure 2 can be obtained from the conventional equivalent circuit (figure 1) by using a referral ratio $a = L_m/L_r'$ in lieu of the common choice of the stator to rotor turns ratio (Novotny & Lipo 1985). With adoption of this ratio, the stator current is seen to be subdivided into the orthogonal components $I_{s\psi}$ (flux component) and I_{sT} (torque component), equivalent to i_{ds}^e and i_{qs}^e referred in the dynamic model, and the slip relation (9) can be obtained by equating the voltages across the parallel branches as

$$\omega_{sl} = s\omega_e = \frac{R_r'}{L_r'}\frac{I_{sT}}{I_{s\psi}}. \tag{17}$$

Equation (17) expresses the co-ordination between the slip and the current components required to attain correct field orientation as relevant for indirect vector control discussed later. The torque expression is obtained from the airgap power as

$$T_e = \frac{3}{2}\frac{P}{2}\frac{L_m^2}{L_r'}I_{s\psi}I_{sT}, \tag{18}$$

which shows the desired torque control via current components $I_{s\psi}$ and I_{sT}.
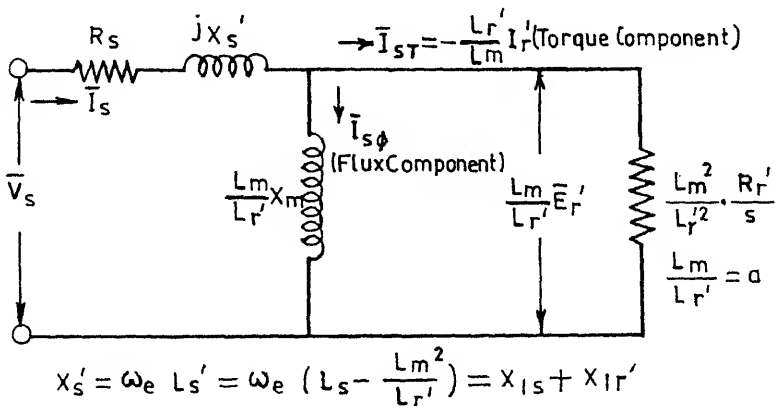
Figure 2. Derived steady state equivalent circuit for rotor flux orientation scheme.

2.3   *Induction motor vector control implementation*

The implementation of vector control requires information regarding the magnitude and the position of the flux vector (stator, rotor or mutual, as the case may be) and fast control of stator current in both magnitude and phase. Depending upon the method of flux acquisition, the vector control can be direct (Blaschke 1972) or indirect (Hasse 1969). The universal field oriented controller developed by De Doncker & Novotny (1988) is applicable to both these field orientation schemes and the generalised approach by Ogaswara *et al* (1988) to indirect control of induction and synchronous motors.

2.3a   *Direct field orientation:*   In the direct method, also known as flux feedback method, the airgap flux is directly measured with the help of sensors such as Hall probes, search coils or tapped stator windings (Zinger *et al* 1990) or estimated/observed from machine terminal variables such as stator voltage, current and speed (Jansen *et al* 1994). Since it is not possible to directly sense rotor flux, it is synthesised from the directly sensed airgap flux using the following equations

$$\psi_{dr}^{s'} = \frac{L_r'}{L_m}\psi_{dm} - L_r' i_{ds}^{s}, \tag{19}$$

$$\psi_{qr}^{s'} = \frac{L_r'}{L_m}\psi_{qm} - L_r' i_{qs}^{s}. \tag{20}$$

A variety of flux observers can be employed to estimate and improve the flux response with less sensitivity to machine parameters as detailed by Verghese & Sanders (1988) and Atkinson *et al* (1991). A major drawback with the direct orientation schemes is their inherent problem at very low speeds when the machine IR drop dominates and the required integration of the signals to measure the airgap flux is difficult. Closed-loop stator flux observers based on the motor current, voltage and the measured rotor position have been found to obviate this difficulty (Jansen *et al* 1993; Lorenz *et al* 1994).

A rotor flux observer based direct vector control scheme as implemented in the laboratory is shown in figure 3 (Chattopadhyay *et al* 1992; Thakur 1996) using a CRPWM inverter with flux and torque regulating loops. The vector rotator block implements the transformation from rotating to stationary axes followed by a 2/3 phase transformation resulting in the following expressions

$$\left.\begin{array}{l} i_a^* = i_{qs}^{*e}\cos\theta_e + i_{ds}^{*e}\sin\theta_e \\[4pt] i_b^* = (-\frac{1}{2}i_{qs}^{*e} - \frac{\sqrt{3}}{2}i_{ds}^{*e})\cos\theta_e + (\frac{\sqrt{3}}{2}i_{qs}^{*e} - \frac{1}{2}i_{ds}^{*e})\sin\theta_e \\[4pt] i_c^* = (-\frac{1}{2}i_{qs}^{*e} + \frac{\sqrt{3}}{2}i_{ds}^{*e})\cos\theta_e - (\frac{\sqrt{3}}{2}i_{qs}^{*e} + \frac{1}{2}i_{ds}^{*e})\sin\theta_e \end{array}\right\}. \tag{21}$$

The speed loop control provides the torque command whereas the flux command is selected according to the operating requirements in either constant torque or constant horsepower region. For CRPWM inverter, line currents are controlled in such a way as to follow the reference current commands generated from the vector rotator.

2.3b   *Indirect vector control:*   An alternative to direct
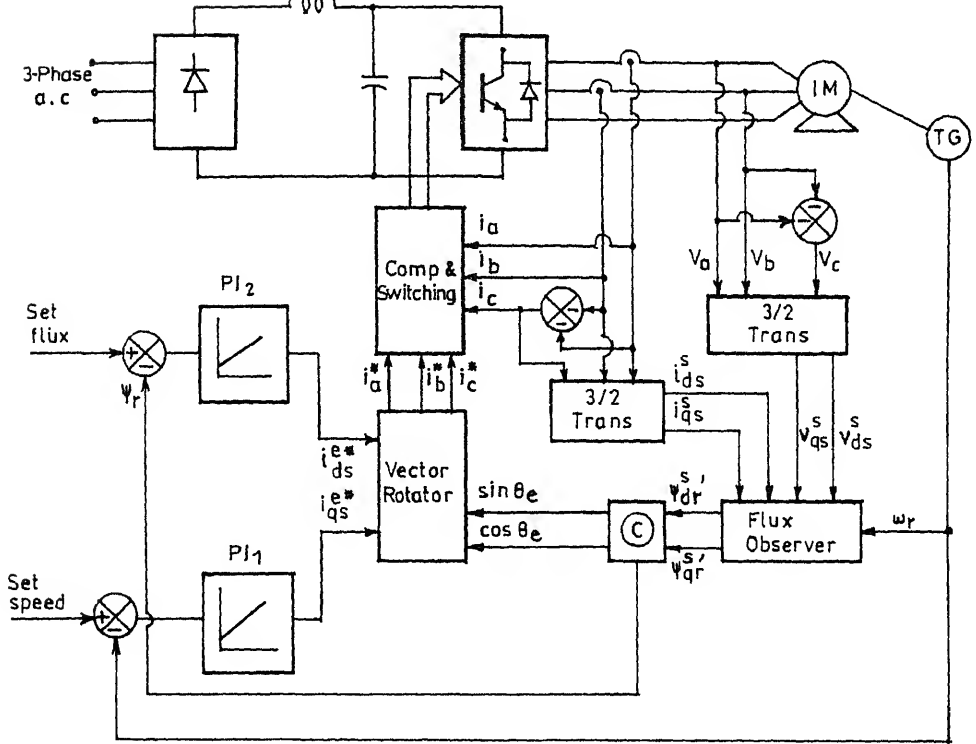
$$\boxed{C} \quad |\hat{\Psi}_r| = \sqrt{(\hat{\Psi}_{dr}^{s\,\prime})^2 + (\hat{\Psi}_{qr}^{\prime})^2}$$

$$\cos\theta_e = \frac{\hat{\Psi}_{dr}^{s\,\prime}}{|\hat{\Psi}_r|} \;,\; \sin\theta_e = \frac{\hat{\Psi}_{qr}^{s\,\prime}}{|\hat{\Psi}_r|}$$

Figure 3. A rotor flux observer based direct vector control scheme for an induction motor with a CRPWM inverter.

to employ the slip relation (9) to compute the flux position relative to the rotor by summing a sensed rotor position signal with a commanded slip position signal

$$\theta_e^* = \theta_{sl}^* + \theta_r. \tag{22}$$

Figure 4 illustrates the basic structure of an indirect field orientation scheme using a CRPWM inverter (Thakur *et al* 1993; Thakur 1996). The commanded currents i_{qs}^{e*} and i_{ds}^{e*} are converted to stator referred reference currents by rotating to stationary and 2/3 phase transformations as in the case of direct field orientation. i_{qs}^{e*} is controlled according to the desired torque and constant rotor flux. i_{ds}^{e*} is obtained from (16) in the steady state.

Indirect field orientation, also known as flux feed-forward control, does not have inherent low speed problems and is preferred in most systems which must have zero speed. However, the inherent limitation is in the slip calculation which depends on the commanded machine parameters that may differ from the actual values during running condition of the drive.
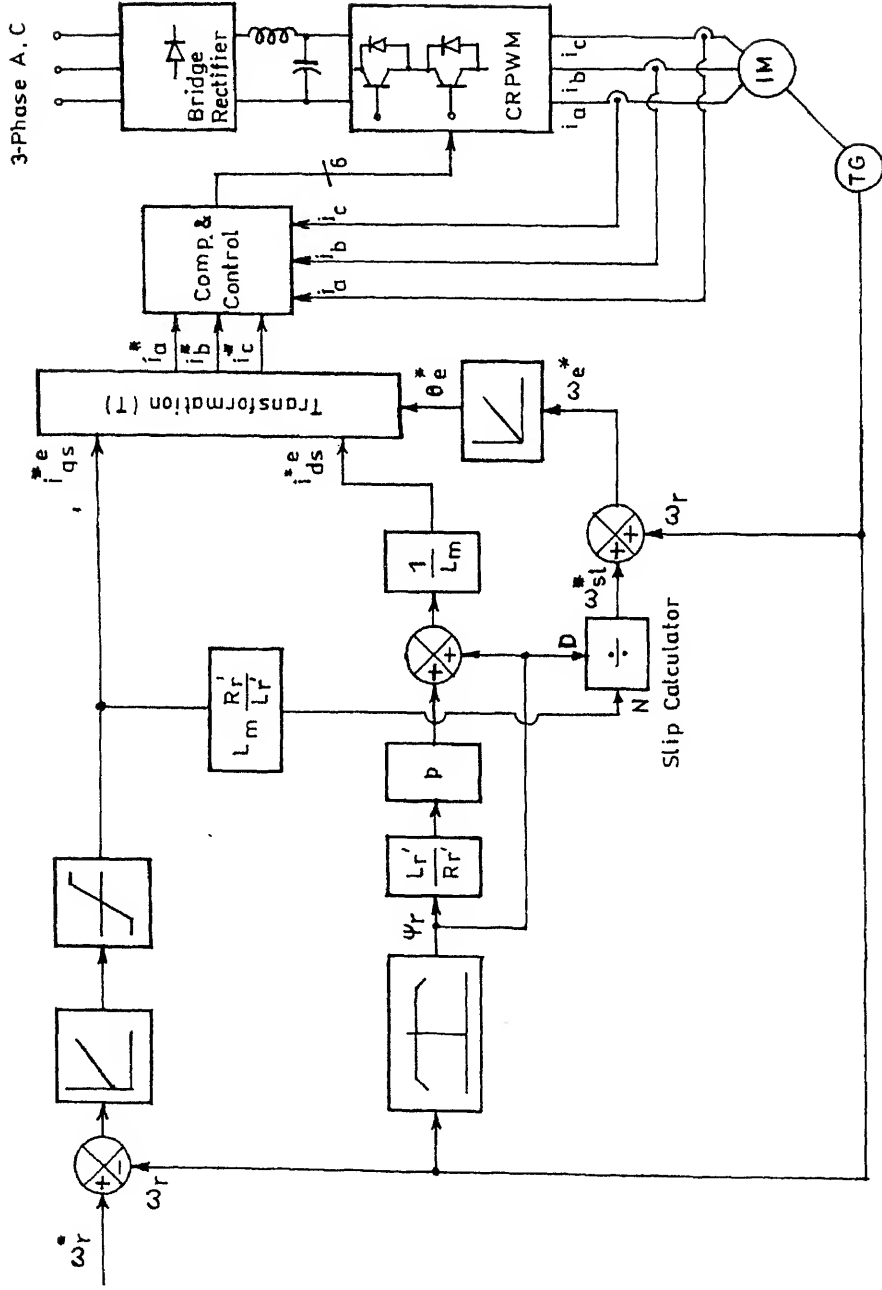
Figure 4. A rotor flux oriented indirect vector control scheme for an induction motor with a CRPWM inverter.
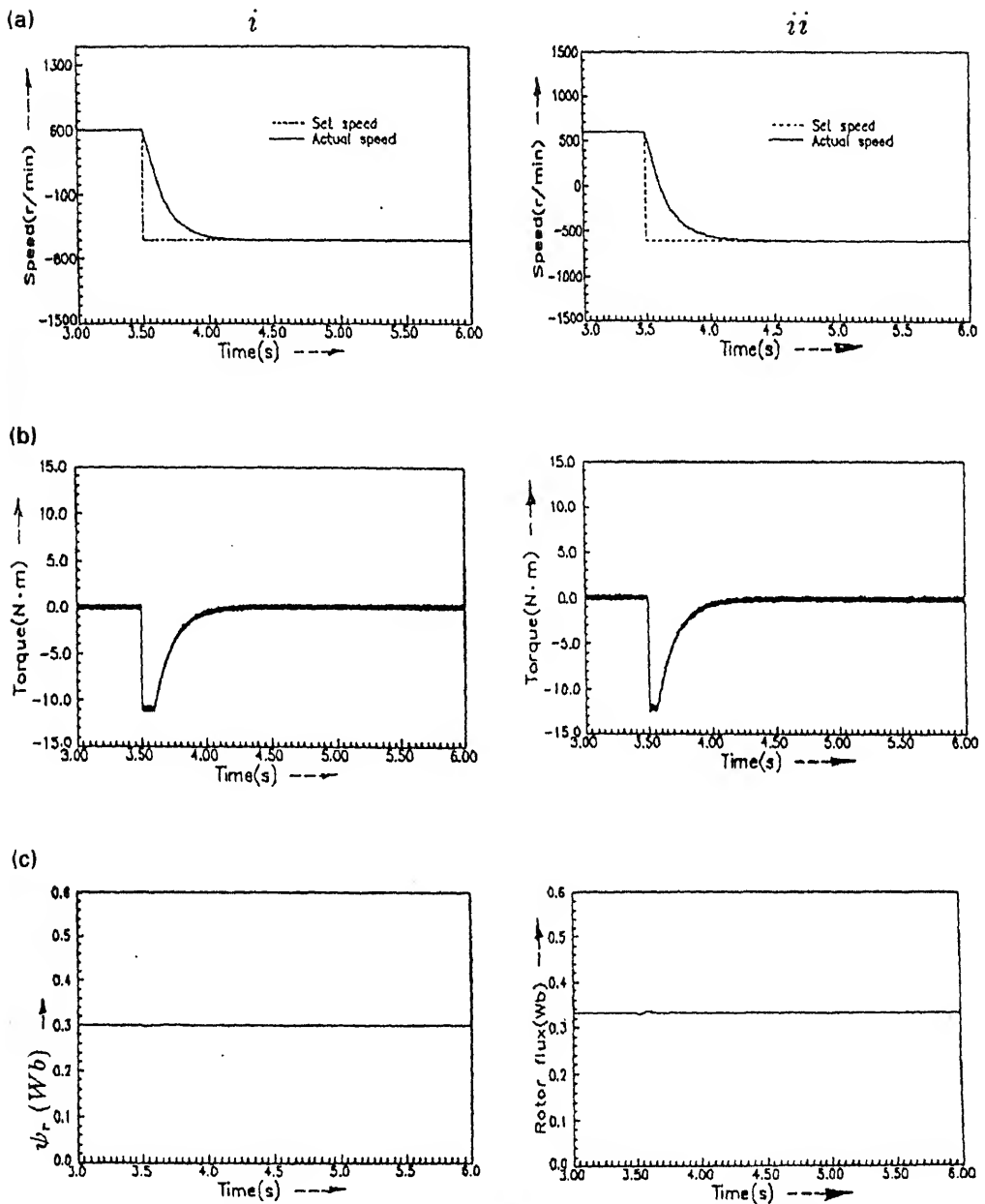
Figure 5. Simulation results showing **(a)** speed **(b)** torque, and **(c)** rotor flux of a vector controlled induction motor drive for speed reversal (600 to −600 rpm): **(i)** direct vector control, **(ii)** indirect vector control.

simpler and cost effective as the differential equations involved are readily solved in real time. Beginning with an 8-bit microprocessor, a 16-bit or a 32-bit or now a DSP, a transputer or a custom-made LSI chip has become a part of the vector control hardware (Gabriel *et al* 1980; Sathiakumar *et al* 1986; Mingbao *et al* 1987; Wu & Strangas 1988; Asher & Sumner 1990; Ho & Sen 1990; Xu & Novotny 1991; Kao & Lin 1992; Lakaparampil 1994).

Multi-microprocessor configuration has also been used to implement sophisticated control structure (Harashima *et al* 1985; Saito *et al* 1987; Tzou & Wu 1990). The limitation in microprocessor application due to its finite word length, execution time and the operational instructions must be taken into account in designing a processor based system as they affect significantly the performance of the system (Dote 1988; Jelassi *et al* 1992). A faster operation may be obtained in a hybrid scheme (Thakur *et al* 1993; Thakur 1996) using both analog hardware and microprocessor based controller where tasks such as 2/3 phase transformation and PWM switching signal generation are achieved with hardware and the PI controller/observer design and implementation by a microprocessor with a PC-XT. Digital computer simulation technique is preferred to optimise the effects of various factors before implementation. Few typical simulation results as obtained and experimentally verified by Thakur (1996) are shown in figure 5 for both indirect and direct vector control. It is seen as expected that the performance of the latter is somewhat superior.

2.4 *Effects of motor parameter variations and adaptation*

Both the schemes described used machine parameters either in the calculation for the slip command for implementing the indirect vector control or to synthesise the flux vector to implement the direct vector control. In the indirect control the main problem is the rotor circuit time constant L'_r / R'_r which is sensitive to both temperature and flux level (Nordin *et al* 1985; Krishnan & Bharadwaj 1991; De Doncker 1994). Direct field orientation systems are sensitive to stator resistance and total leakage inductance but, typically, the parameter sensitivity is less here than that with the indirect control, specially because of the flux regulation through feedback. With deviation of parameters, the field orientation is not perfect and the controller should track the machine parameters. Several methods of parameter adaptation have been attempted (Garces 1980; Matsuo & Lipo 1985; Dalal & Krishnan 1987; Krishnan & Doran 1987; Nilsen & Kazmeirkowski 1989; Bal & Grant 1992; Ghosh & Bhadra 1992), along with a number of identification schemes including Model Reference Adaptive control (Ohnishi *et al* 1986; Holtz & Thimm 1989; Vas 1990; Bal & Grant 1992; Moriera & Lipo 1993). Automated initial tuning in the form of self-commissioning technologies has also been developed (Khambadkone & Holtz 1991; Lorenz *et al* 1994; Borgard *et al* 1995; Yanagawa *et al* 1995). Recent work on the on-line tuning to improve the robustness of vector control induction motor has been reported using special torque control strategy (Noguchi *et al* 1997; Tadakuma *et al* 1997) and feed forward/feedback control with neural network. A new flux and stator resistance identifier for *ac* drives has been proposed by Kerkman *et al* (1996).

Two new approaches to induction motor field-orientation are presented in Matsuo *et al* (1994) which employ rotor end ring current phase detection to make the controller

independent of rotor time constant variations. However, it has been reported that the control performance is adequate within the normal operating temperature for most of the high performance applications and the parameter adaptation may be essential only in the case of critical applications. The parameter sensitivity in small machines is low enough to cause serious problems (Nordin *et al* 1985).

Issues regarding field-oriented controller for induction motors with double cage and deep bar rotor are discussed in Vas (1990). For these motors, the angular slip torque has to be calculated in such a way that it contains the effects of the deep bar or the double cage. Improved cage rotor models are developed by Healey *et al* (1995).

2.5 Effects of magnetic saturation and core loss

The flux level in an induction machine is a function of both the stator and the rotor currents. Both the performance and the losses are effected by its selection (Khater *et al* 1987). Normal modelling of the machine will not remain valid under magnetic saturation, particularly so under dynamic condition. The saturation effects for vector-controlled machines have been considered by Lorenz & Novotny (1990) and Vas & Alakul (1990). Under saturation conditions, the peak torque per ampere is best produced by increasing the torque producing current command in proportion to the total stator current. The sensitivity of rotor flux estimation depends on the selection of the machine model (Levi & Vuckovic 1989, 1990). The load torque condition has been observed to play an important role in machine saturation (Ohm 1989).

Vector control principles have been traditionally derived on the assumption that the iron core loss may be neglected. However, recently, it has been shown (Levi 1995; Levi *et al* 1996) that the core loss introduces unwanted cross coupling leading to detuning and for compensation, a decoupling circuit for indirect rotor flux oriented control is suggested, which makes the controller more complex.

2.6 Current, flux and torque regulators

Current regulators for vector controlled *ac* drives are more complex than those for *dc* drives as both amplitude and phase of the stator current are to be controlled. Both CSI and PWM converters with current regulation are used. The current regulators classified into three groups, hysteresis, PI with ramp comparison PWM and predictive (optimal) voltage vector location have been adequately discussed by Lorenz *et al* (1994) and Lee *et al* (1994). The various solutions differ in implementation costs, robustness with respect to parameter variation and their ability to track current commands with high fidelity and low distortion.

Regulation of flux is limited by the estimation of the flux magnitude and angle in direct vector control. Both open loop and closed loop flux observers have been used for direct and indirect field orientation (Hillenbrand 1977; Bouch *et al* 1992; Jansen *et al* 1993; Lorenz *et al* 1994). It was shown that a position sensor along with a current sensor will facilitate a simple open-loop observer for rotor flux. The closed-loop observer with motor current, voltage and rotor position measurement using the best features of both the current model and the voltage model open-loop observers will give better performance – flux regulation as well as flux estimation.

Recently, fuzzy and neural network-based estimators of feedback signals such as rotor flux, unit vectors and torque for indirect and direct vector control schemes have been reported (Miki *et al* 1991; Sousa & Bose 1993; Simoes & Bose 1995). These have the advantages of faster execution speed, harmonic ripple immunity and fault tolerance characteristics compared to a DSP based estimation.

2.7   *Direct torque control (DTC)*

The latest control method developed and commercialised by ABB, Sweden from the concept of the field-oriented or vector control is direct torque control (DTC), a patented concept developed again in Germany by Depenbrock (1988). The basic control scheme is shown in figure 6 when both the flux and torque are controlled by a hysteresis controller (Tiitinen *et al* 1996; Nash 1997). The delays associated with the PWM stage are eliminated since the PWM modulation is replaced by an optimal switching (Space PWM) logic. The adaptive motor model estimates the actual torque, stator flux and shaft speed as well as the frequency. The flux and torque are calculated every 25 μs and the speed and the frequency once per millisecond. The input to the motor model includes the motor current for two stator phases, line voltage and power switch positions. The optimal switching logic is realised by ASIC hardware (ACS 600). The switch information for the power module is utilised in the calculation of the appropriate voltage vector which will satisfy both the torque status and flux status outputs. This method results in a better torque response than the flux vector control and, in addition, assuming moderate speed accuracy is acceptable (typically 0.1–0.3%), the need for a pulse encoder is eliminated. Implementations of special functions
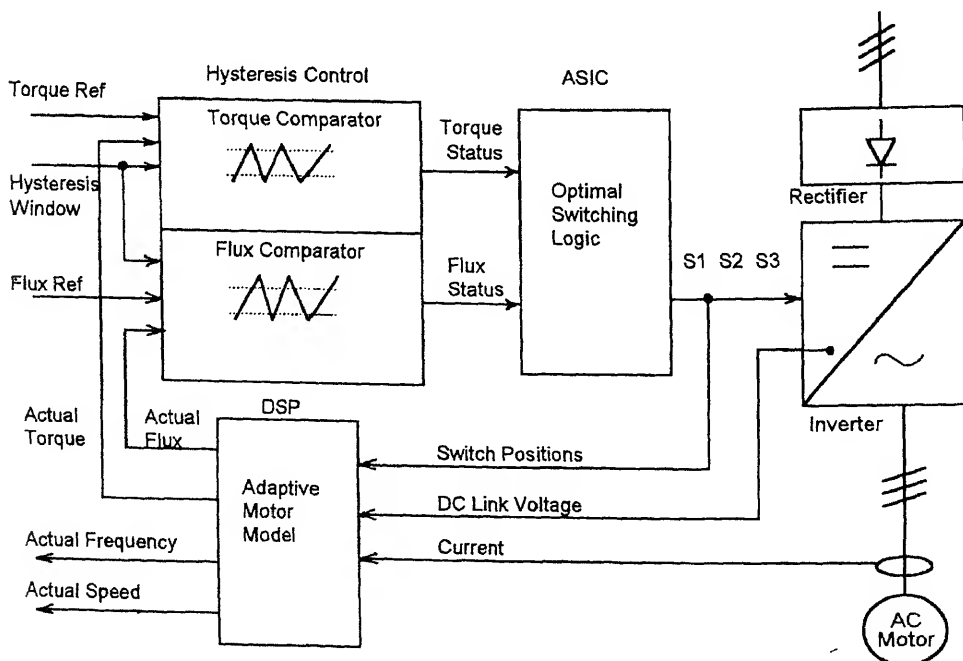
Figure 6.   Direct torque control (DTC) scheme.

flying start, flux braking, flux optimisation and powerloss ride through are all made
er with this control approach, as claimed.

Doubly-fed and multiphase induction motor control

vector control of a doubly-fed slip ring induction motor in a Scherbius scheme as used
igh power pump drives with a current controlled cycloconverter in the rotor side is
ly described by Vas (1990) and Bose (1986) for super/sub-synchronous speed control.
same system can be used for VSCF generation systems, where the control strategy
ains the same except that the active and the reactive currents of the cycloconverter are
trolled to control real and reactive powers, respectively, at the stator terminals by the
lback method. A novel control strategy to realise torque and reactive power control of
ubly excited induction machine with position sensorless scheme using rotor voltage
currents as feedback signals has been proposed recently by Xu & Cheng (1995).
strategy for improvement of the reliability for vector-controlled induction motor drive
a modified topology where the neutral point is returned to the midpoint on the *dc* link
roposed by Liu *et al* (1993). This allows for continuous disturbance-free operation of
drive even with complete loss of one leg of the inverter or motor phase. This method
been extended to field-oriented control for a multiphase induction machine with an
alanced stator winding structure (Zhao & Lipo 1996)

Vector control of synchronous motors

Brief application survey

ile vector-controlled induction motor drives have been used mostly in the industry for
lium power ranges, vector-controlled synchronous motor drives are either in the very
power range (1–10 MW) with wound-field machines fed from cycloconverters or in the
kilowatt range with permanent magnet synchronous motors (PMSM) or synchronous
ctance motors for servo drives. The control of synchronous motors is different from
of induction motors primarily due to the fact that in the former, the magnetising current
be supplied from the field side independently of the armature current and the space
ition of the field is located by the position of the rotor. Additionally, the steady state
between the rotor (which usually carries the field winding) and the controlled flux
tor vanishes in the steady state. Therefore, the indirect or the feed-forward type of
tor control as used extensively for the induction motor drives does not apparently seem
ious for a synchronous machine. The 'transvector control' as applied to a synchronous
tor by Bayer *et al* (1972) is essentially a direct type of flux feedback control where the
or current is orthogonally oriented with respect to the stator flux vector to achieve unity
dy state power factor. The decoupling is achieved by a closed loop flux feedback in
ition to feeding a part of the magnetising current from the stator during the transient
compensate for a sluggish field current change. Siemens has reported (Timpe 1982;
lmann 1992) the development of vector controlled cycloconverter-fed subsynchronous
tor drives for use in reversing rolling mills to achieve high dynamical control response.
wn Boveri reported the development of the first gearless tube mill (Blauenstein 1970;

Stemmler 1970) using flux feed-forward control scheme. Terens *et al* (1982) used both static and dynamic flux models to control a similar drive. Nakano *et al* (1984) reported the development of a high performance synchronous motor drive for a rolling mill, with an open-loop flux estimator and PI current controller. An airgap flux oriented vector controlled cycloconverter drive was developed by Hill *et al* (1987) for an icebreaker. A very good survey of field-oriented control of synchronous machines including various applications has been made by Novotny & Jansen (1991) with a discussion on the difference between the 'space angle control' (SAC) relevant to self-synchronous commutatorless motor (CLM) and the true field-oriented (FO) or vector control. While in the former the angle of the armature current vector with respect to the field axis may be other than 90°, in the latter it is strictly restricted to 90°. Earlier, high power drives using CSI converters and wound field-synchronous motors with load commutation for fan and compressor drives were SAC systems utilising a rotor position detector to cause the power converter to supply stator excitation in synchronism with the induced voltage from the field excitation. PMSM motors operated in true FO system as used for servo drives and machine tools are reported by Kaufman *et al* (1982) and Wescheta (1983).

3.2 Synchronous machine model and vector control implementation

3.2a *Wound field synchronous motor model:* The d-q model of a wound field salient pole synchronous machine with damper windings in Park (rotor) reference frame is

$$
\begin{bmatrix} v_{qs} \\ v_{ds} \\ 0 \\ 0 \\ v'_{fr} \end{bmatrix} = \begin{bmatrix} R_s + pL_{qs} & \omega_r L_{ds} & pL_{qm} & \omega_r L_{dm} & \omega_r L_{dm} \\ -\omega_r L_{qs} & R_s + pL_{ds} & -\omega_r L_{qm} & pL_{dm} & pL_{dm} \\ pL_{qm} & 0 & R'_{qr} + L'_{qr} & 0 & 0 \\ 0 & pL_{dm} & 0 & R'_{dr} + pL'_{dr} & pL_{dm} \\ 0 & pL_{dm} & 0 & pL_{dm} & R'_{fr} \\ & & & & +p(L'_{lfr} + L_{dm}) \end{bmatrix} \begin{bmatrix} i_{qs} \\ i_{ds} \\ i'_{qr} \\ i'_{dr} \\ i'_{fr} \end{bmatrix},
$$

(23)

$$
T_e = \frac{3}{2}\frac{P}{2}(\psi_{ds}i_{qs} - \psi_{qs}i_{ds})
$$

(24)

$$
= T_L + \frac{2}{P}J\frac{d\omega_r}{dt}.
$$

(25)

The field circuit parameters $v'_{fr}, i'_{fr}, R'_{fr}, L'_{lfr}$ refer to the stator. For steady state, p-terms vanish,

$$
i_{qs} = I_{qs}, \quad i_{ds} = I_{ds}, \quad i'_{dr} = i'_{qr} = 0, \quad i'_{fr} = I_f.
$$

3.2b *Vector control and angle control:* The rotor position feedback and vector control of the motor stator current to maintain the space angle between the field winding and the stator mmf results in stator currents that translate to set values of i_{qs} and i_{ds} in the rotor reference frame. This is due to the instantenenous control of the phase of the stator current to always maintain the same orientation of the stator mmf vector with respect to the field winding in the d-axis of the d-q model. The resulting axes current are shown in
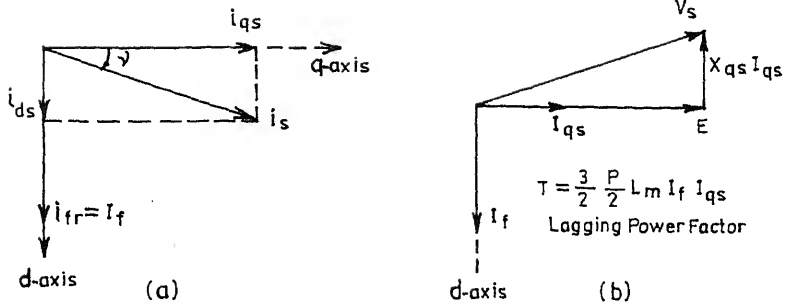
Figure 7. Phasor diagram of synchronous machine currents in d-q axes: (a) space angle control, (b) field orientation ($\gamma = 0$, $I_{ds} = 0$).

figures 7a and b (Novotny & Lipo 1985) for space angle control and the field orientation (when $\gamma = 0$, $I_{ds} = 0$). Note that for field orientation the field current in the d-axis and the stator current in the q-axis are 90° apart.

3.2c *Implementation with CSI and CRPWM inverter:* The implementation calls for control of magnitude and phase of the stator current with respect to the location of the field winding axis. Figure 8 shows a direct implementation ($\gamma = 0$) using absolute rotor position sensing and a CSI. With $\gamma = 0$, the stator current is entirely q-axis current and is equivalent to a torque command. The γ^* command is entered in the 'phase regulator' block and the drive can be operated at other than $\gamma^* = 0$. Figure 9 shows a simple means for implementing torque control with independent q-axis and d-axis currents using a CRPWM. The absolute rotor position information is used to convert the i_{qs}^{e*} and i_{ds}^{e*} commands in the rotor reference frame to a stator reference frame – which become the current commands for the CRPWM. Normal field orientation is obtained by setting $i_{ds}^{e*} = 0$. Varying i_{ds}^{e*} provides control of power factor and other varying performances. The 'rotor to stator transformation' block in figure 9 implements the same equations as in (21).

3.2d *Implementation with cycloconverter:* Cycloconverter-fed synchronous motors have been preferred for low speed large power drives e.g. mine hoist winders, gearless
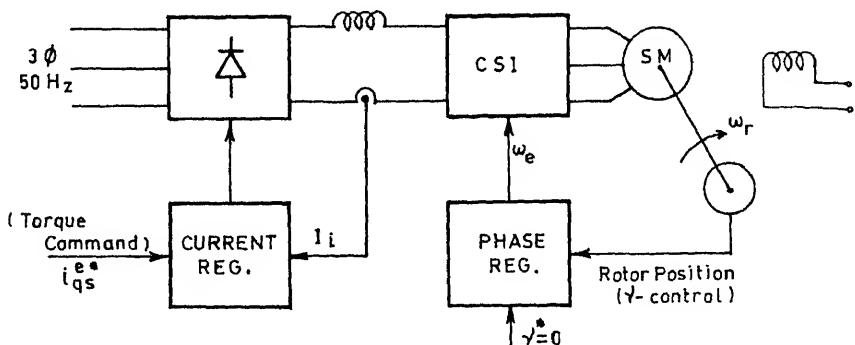
Figure 8. A synchronous motor vector control scheme using a CSI ($\gamma = 0$, $I_{ds}^{*e} = 0$).
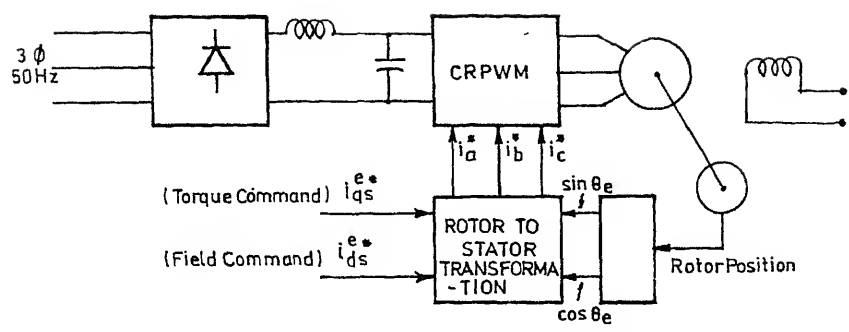
Figure 9. A synchronous motor vector control scheme using a CRPWM ($\gamma = 0$, $I_{ds}^{*e} = 0$).

6PCC: 6-Pulse Cycloconverter & Control

VR: Vector Rotator
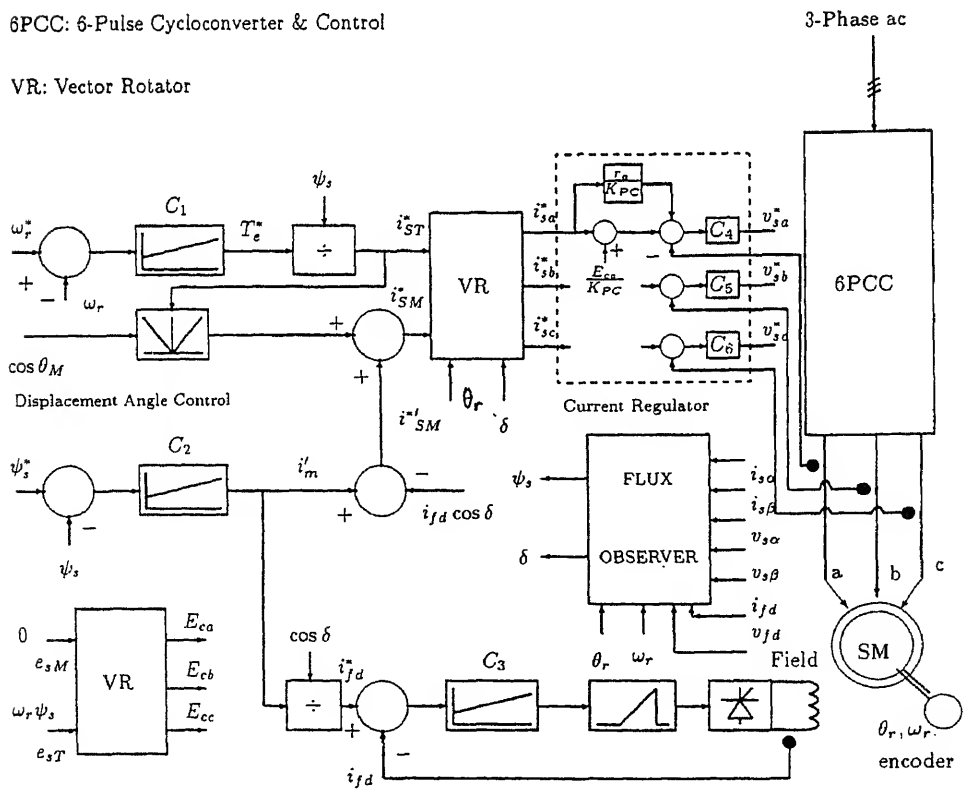
Figure 10. A stator flux oriented vector control scheme for a cycloconverter fed synchronous motor with a flux observer.

nent mill drives, rolling mill drives, ship propulsion drives etc. Synchronous motors
e been preferred in these drives rather than induction motors because of their power
tor and large torque capability at low speeds. Furthermore, a naturally commutated
loconverter, compared to an inverter, provides a near-sinusoidal current excitation re-
ting in negligible torque ripple, inherent four quadrant capability, robustness and large
wer handling capability.

A stator flux-oriented vector control scheme which is an improvement of that in Bayer
al (1972) and Nakano *et al* (1984) for a 6-pulse non-circulating current cycloconverter-
synchronous drive with a flux observer has been developed recently by Das (1996)
d Das & Chattopadhyay (1997) for a rolling mill drive. Figure 10 shows the implemen-
on scheme which aims at a control that maintains a spatial orthogonality between the
x vector ψ_s and the armature current vector i_a as shown in the space phasor diagram
figure 11. The reference speed and reference flux commands are given to the vector
ntroller that generates the reference analog voltages for the cycloconverter (through the
rrent controller) and the field converter. The stator flux is estimated by a closed-loop
duced order observer. Referring to figure 10, C_1 is the speed controller that generates
torque command which is divided by the stator flux to generate the torque command
current i_{sT} . The magnetisation current along the flux axis (i'_m) is obtained from a flux
ntroller C_2 . The transient stator flux component of current i_{sm} is obtained from the re-
ionship, $i^{*'}_{sm} = i'_m - i_{fd} \cos \delta$, which decays down to zero in the steady state. The steady
te displacement angle is decided by the displacement angle controller. The set value
the field current is obtained from the relation, $i^*_{fd} = i'_M / \cos \delta$. C_3 is the field current
ntroller that generates the control voltage for triggering the field converter. The vector
ator (VR) transforms the vector from two axes flux – torque reference frame to abc
tionary reference frame. The observer and the control circuit design aspects together
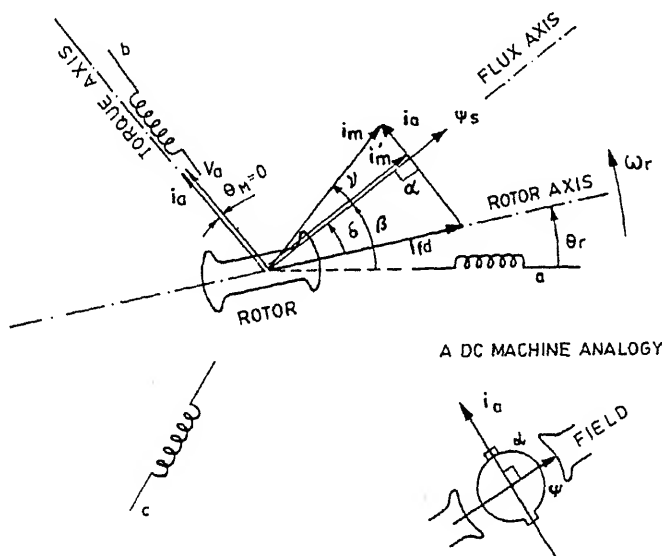
Figure 11. Space phasor diagram for vector controlled synchronous motor.
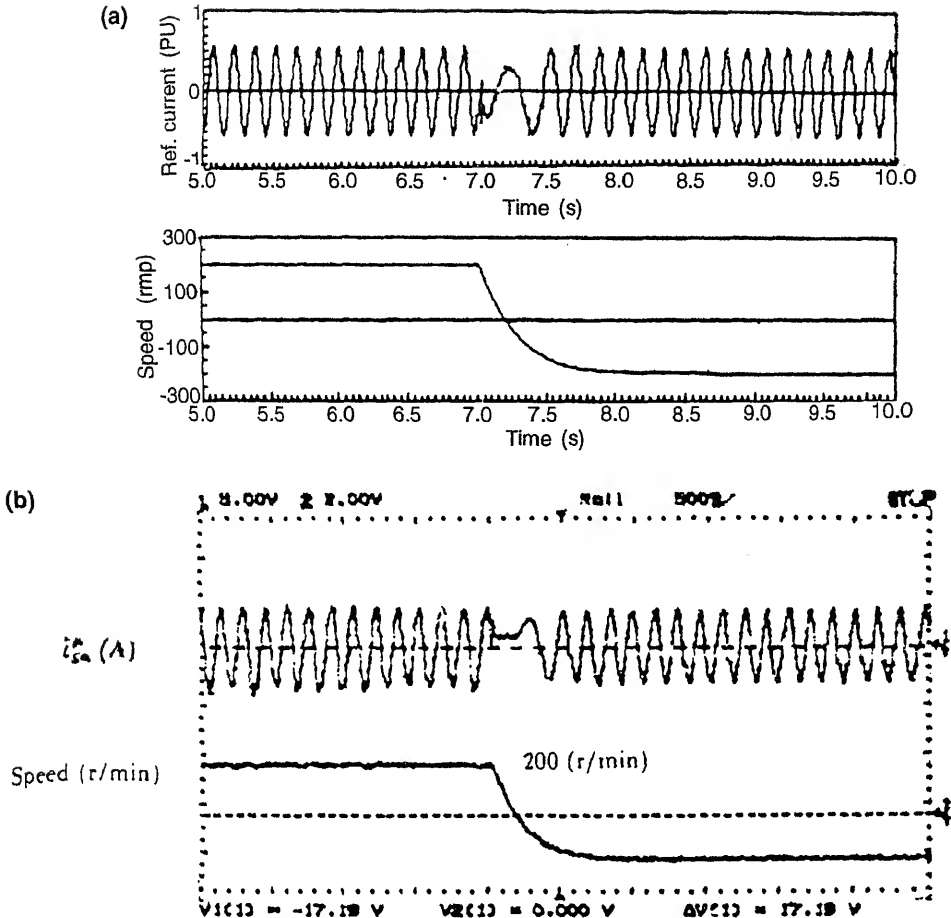
Figure 12. Speed and current responses to a step reversal (200 to −200 rpm) of a vector controlled cycloconverter fed synchronous motor drive: (a) simulation results, (b) experimental results.

with the PC-based implementation are detailed in Das (1996) and Das & Chattopadhyay (1997) and typical results obtained are shown in figure 12.

Vector control of a synchronous motor can be made with respect to three flux/mmf vectors, namely, the stator flux, the damper flux and the field mmf. A unified analysis of these three schemes by Das (1996) shows that the stator flux orientation results in a unity power factor which is not the case with the other schemes. A damper flux orientation scheme which is comparable to rotor flux orientation in the induction motor drive has been recently reported by Chongjium *et al* (1995) without any detailed analysis. Orientation with field mmf results in a lagging motor terminal power factor and is not expedient for high power drives.

3.2e *Saturation and damper effects:* Magnetic saturation effects of the d-axis and q-axis for a damperless, salient pole stator flux-oriented wound rotor synchronous motor drive have been studied by Brass & Mecrow (1992) by developing a saturated flux model. In a separate paper (Brass & Mecrow 1993), the effect of damper windings on field-oriented
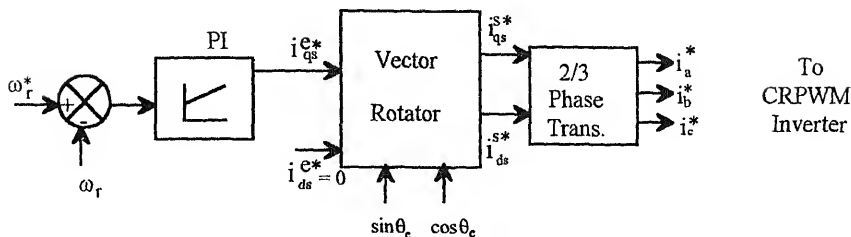
Figure 13.   Vector control scheme for a permanent magnet synchronous motor.

synchronous motor has been studied and it is shown that the presence of damper windings improves the torque dynamics.

3.2f   *Vector control of permanent magnet synchronous and reluctance motors:*   Permanent magnet machines, both surface-mounted and interior-magnet types, are extensively used in servo drives and robotic applications, and vector control provides smooth torque operation of these motors through the entire speed range, including zero, with high power factor. The control schemes for these machines have been extensively discussed in Vas (1990) and so are not detailed here. In a permanent magnet synchronous motor, the rotor field flux ψ_f and the corresponding equivalent field current I_f can be considered as constant. For surface-mounted machines, the saliency and armature reaction is negligible. Therefore, $\psi_f = \psi_m$ and for maximum torque sensitivity with stator current $I_{ds}^{e*} = 0$, and $I_s^* = i_{qs}^e$. Figure 13 shows the vector control principle for the PMSM derived from the induction motor control diagram with the modifications (Bose 1986), $w_{sl} = 0$, $\theta_r = \theta_e$. A microprocessor-based field-oriented control scheme for a permanent magnet hysteresis synchronous motor is presented in Qian & Rahaman (1993).

Synchronous reluctance motor drives have recently received renewed attention due to the application of field oriented control to these motors (Boldea *et al* 1991; Xu *et al* 1991; Matsuo & Lipo 1993). Excellent control performance of the drive systems has been obtained though there exists a limitation in the field weakening range.

4.   Conclusions

The vector control of *ac* drives in which there have been a spurt of activities, has by now gained maturity but still continues to provide interesting and challenging scope for innovations to researchers and application engineers. This paper has made an attempt to make a summary review of the activities on various aspects in this important field for control of both induction and synchronous machines till date with the informations available in the published literature. It is expected that the review will help those interested in the development of efficient and high performance drives of the future.

References

Asher G M, Sumner M 1990 Parallelism and transputer for real time control of *ac* induction motors. *IEE Proc.* D137: 179–188

Atkinson D J, Acernly P P, Finch J W 1991 Observers for induction motor state and parameter estimation. *IEEE Trans. Ind. Appl.* 27: 1119–1127

Bal G, Grant D M 1992 Parameter estimation of field oriented controlled induction motor fed by CRPWM via EKF using stator reference currents. *Conf. Proc. Int. Conf. Elec. Machines* (Manchester), pp 597–601

Bayer K H, Blaschke F 1977 Stability problem with the control of induction motors using method of field orientation. *Conf. Rec. Int. Fed. Autom. Control* (Lussane), pp 483–492

Bayer K H, Waldmann H, Weibelzahl 1972 Field oriented close-loop control of a synchronous machine with the NEW transvector control system. *Siemens Rev.* 39: 220–223

Blaschke F 1972 The principle of field orientation as applied to the new TRANSVECTOR closed loop control system for rotating field machines. *Siemens Rev.* 39: 217–220

Blauenstein 1970 The first gearless drive for a tube mill. *Brown Boveri Rev.* 57: 96–105

Boldea I, Nasar S A 1992 *Vector control of AC drives* (West Palm Beach, FL: CRC)

Boldea I, Fu Z X, Nasar S A 1991 Torque vector control (TVC) of axially laminated anisotropic (ALA) rotor reluctance synchronous motors. *Elec. Mach. Power Syst.* 19: 381–398

Borgard D E, Olsson G, Lorenz R D 1995 Accuracy issues for parameter estimation of field oriented induction motor drives. *IEEE Trans. Ind. Appl.* 31: 795–801

Bose B K 1986 *Power electronics and ac drives* (Englewood Cliffs, NJ: Prentice Hall)

Bouch H, Blumel R, Zeng W 1992 Flux estimation of a PWM inverter fed torque controlled induction machine based on terminal quantities. *Conf. Proc. Int. Conf. Elec. Machines* (Manchester), pp 833–837

Brass M A, Mecrow B C 1992 The accommodation of saturation in the control of field oriented synchronous drives. *Conf. Rec. Int. Conf. Elec. Machines 92* (Manchester), pp 848–852

Brass M A, Mecrow B C 1993 The role of damper circuits in field oriented synchronous motors. *Proc. Int. Elec. Eng. Conf.* (London), pp 115–120

Chattopadhyay A K, De N K, Thakur A N 1992 Studies on a rotor flux observer based direct type vector controlled induction motor. *Conf. Rec. Int. Conf. Elec. Rotating Machines* (Bombay) 2A: 1–8

Chongjium L, Chunyi Z, Yaohua L, Jingde G 1995 A high performance synchronous motor field oriented system. *Conf. Rec. IEEE* (Singapore), pp 825–827

Dalal D, Krishnan R 1987 Parameter compensation of indirect vector controlled induction motor drive using estimated airgap power. *IEEE Ind. Appl. Soc. Annu. Meet Conf. Rec.*, pp 170–176

Das S P 1996 *Design, simulation and PC-based implementation of a high performance cyclo-converter fed synchronous motor drive system.* Ph D thesis, Indian Institute of Technology, Kharagpur

Das S P, Chattopadhyay A K 1997 Observer based stator flux oriented vector control of cycloconverter-fed synchronous motor drive. *IEEE Trans. Ind. Appl.* 33: 943–955

De Doncker R W 1994 Parameter sensitivity of indirect universal field oriented controller. *IEEE Trans. Power Electron.* 9: 367–375

De Doncker R W, Novotny D W 1988 The universal field oriented controller. *Conf. Rec. IEEE Ind. Appl. Soc. Annu. Meet*: 450–456 (also 1994 *IEEE Trans. Ind. Appl.* 30: 92–100

Depenbrock M 1988 Direct self control (DSC) of inverter fed induction machine. *IEEE Trans. Power Electron.* 3: 420–429

Dote Y 1988 Application of modern control technology to motion control. *Proc. IEEE* 76: 438–454

Erdman W L, Hoft R G 1990 Induction machine field orientation along airgap and stator flux. *IEEE Trans. Energy Conversion* 5: 115–121

briel R, Leonhard W, Nordby C J 1980 Field orientation control of standard *ac* motors using microprocessor. *IEEE Trans. Ind. Appl.* 16: 186–192

ces L J 1980 Parameter adaptation for speed controlled static *ac* drive with squirrel cage nduction motor. *IEEE Trans. Ind. Appl.* 16: 173–178

osh B C, Bhadra S N 1992 DC link voltage based rotor resistance adaptation scheme of a field oriented CSI-IM drive system. *Conf. Rec. Int. Conf. Elec. Rotating Machines* (Bombay) 2A: 15–22

ashima F, Kondo S, Ohnishi K, Kajita M, Susono M 1985 Multi microprocessor- based control system for quick response induction drive. *IEEE Trans. Ind. Appl.* 21: 602–609

sse K 1969 *Zur dynamik drehzahlgeregelter antriebe mit stromrichtergespeisten asynchron-kurzschlufermaschinen*. Ph D dissertation, Tech. Hochschule, Darmstadt

sse K 1977 Control of cycloconverter for feeding of asynchronous machines. *Conf. Rec. IFAC* (Lussane) : 537–545

aley R C, Williamson S, Smith A C 1995 Improved cage rotor models for vector controlled induction motors. *IEEE Trans. Ind. Appl.* 31: 812–822

l W A, Turton R A, Dungen R J, Schwalm C L 1987 A vector controlled cycloconverter drive for an icebreaker. *IEEE Trans. Ind. Appl.* 23: 1036–1041

lenbrand F 1977 A method of determining the speed and rotor flux of the asynchronous machine by measuring the transient quantities. *Conf. Proc. IFAC Control of Power Electronics and Electrical Drives* (Lussane), pp 55–62

E E Y, Sen P C 1988 Decoupling control of induction motor drives. *Proc. IEEE* 35: 253–262

E Y Y, Sen P C 1990 A microcontroller-based induction motor drive system using variable structure strategy with decoupling. *IEEE Trans. Ind. Electron.* 37: 227–235

ltz J, Thimm T 1989 Identification of machine parameters in a vector controlled induction motor drive. *Conf. Rec. IEEE Ind. Appl. Soc. Annu. Meet.* 601–606

ang L, Tadokorro Y, Matsuse K 1994 Deadbeat flux level control of direct field oriented high horsepower induction servo motor using adaptive rotor flux observer. *IEEE Trans. Ind. Appl.* 30: 954–962

sen P L, Lorenz R D, Novotny D W 1993 Observer based direct field orientation : analysis and comparison of alternative methods. *Conf. Rec. IEEE Ind. Appl. Soc. Annu. Meet* (Toronto), pp 536–543

sen P L, Lorenz R D, Novotny D W 1994 Observer based direct field orientation: analysis and compensation of alternating methods. *IEEE Trans. Ind. Appl.* 30: 945–953

assi K, Fornel D, David M P 1992 Numerical considerations in field oriented control of asyn-chronous drives. *Conf. Proc. Int. Conf. Elec. Machines* (Manchester), pp 632–636

o Y T, Lin C H 1992 Analysis and design of microprocessor based vector controlled induction motor drives. *IEEE Trans. Ind. Electron.* 39: 96–105

ufman G, Garces L, Gallaghar 1982 High performance servo drives for machine tool applica-tions using *ac* motors. *IEEE Ind. Appl. Soc. Conf. Rec.*, pp 604–609

rkman R J, Seibal B J, Rowan T M, Schlegal D W 1996 A new flux and stator resistance identifier for *ac* drive system. *IEEE Trans. Ind. Appl.* 32: 585–593

ambadkone A M, Holtz J 1991 Vector controlled induction motor drive with a self commis-sioning scheme. *IEEE Trans. Ind. Electron.* 38: 322–327

ater F, Lorenz R D, Novotiny D W, Tang K 1987 Selection of flux level in field oriented induction machine controllers with consideration of magnetic saturation effects. *IEEE Trans. Ind. Appl.* 23: 276–282

ishnan R, Bharadwaj A S 1991 A review of parameter sensitivity and adaptation in indirect

Krishnan R, Doran F C 1984 Study of parameter sensitivity in high performance inverter fed induction motor drive systems. *IEEE Ind. Appl. Soc. Annu. Meet* : 510–514

Kume T, Iwakane T 1987 High performance vector controlled *ac* motor drives: application and new technologies. *IEEE Trans. Ind. Appl.* 23: 872–880

Lakaparampil Z V 1994 *Digital controllers for high power and high performance induction motor drives.* Ph D thesis, Indian Institute of Science, Bangalore

Lee D C, Sul S K, Park M H 1994 High performance current regulator for field oriented controlled induction motor drive. *IEEE Trans. Ind. Appl.* 30: 1247–1257

Leonhard W 1985 *Control of electric drives* (Berlin: Springer-Verlag)

Leonhard W 1986 Microprocessor control of high performance *ac* drive-a survey. *Automatica* 22: 1–19

Levi E 1995 Impact of iron loss on behaviour of vector controlled induction motors. *IEEE Trans. Ind. Appl.* 31: 1287–1296

Levi E, Vuckovic V 1989 Field oriented control of induction machines in the presence of magnetic saturation. *J. Elec. Mach. Power Syst.* 16: 133–147

Levi E, Vuckovic V 1990 A method of rotor flux estimation in saturated field oriented induction machines. *Conf. Rec. Int. Conf. Elec. Machines*, pp 344–349

Levi E, Sokola M, Boglielli A, Pastorelli M 1996 Iron loss in rotor flux oriented induction machines: identification, assessment of detuning and compensation. *IEEE Trans. Power Electron.* 11: 698–709

Liu T H, Fu J F, Lipo T A 1993 A strategy for improving reliability of field oriented controlled induction motor drives. *IEEE Trans. Ind. Appl.* 29: 910–918

Lorenz R D, Novotny D W 1990 Saturation effects in field oriented induction machines. *IEEE Trans. Ind. Appl.* 26: 283–290

Lorentz R D, Lipo T A, Novotny D W 1994 Motion control with induction motors *Proc. IEEE* 82: 1215–1240

Matamo T, Blasko V, Moreira J C, Lipo T A 1994 Field oriented control of induction machines employing rotor end ring current detection. *IEEE Trans. Power Electron.* 9: 638–645

Matsuo T, Lipo T A 1985 A rotor parameter identification scheme for vector controlled induction motor drives. *IEEE Trans. Ind. Appl.* 21: 624–632

Matsuo T, Lipo T A 1993 Field oriented control of synchronous reluctance machine. *IEEE Conf. Rec. Power Electron. Syst. Conf.* : 425–431

Miki I, Nagai N, Sakae N, Yamada T 1991 Vector control of induction motor with fuzzy PI controller. *IEEE/Ind. Appl. Soc. Annu. Meet Conf. Rec.*, pp 342–346

Mingbao Z, Wenlong Q, Heping Z, Bring H 1987 An adjustable speed three phase motion control by a Z-80 single board micro-computer using vector control. *Conf. Proc. Evolution and Modern Control of Induction Machines* (Torino) : 513–518

Moriera J C, Lipo T A 1993 A new method for rotor time constant tuning in indirect field oriented control. *IEEE Trans. Power Electron.* 8: 626–631

Murphy J M D, Turnbull F G 1988 *Power electronic control of ac motors* (New York: Pergamon)

Nakano T, Ohsawa H, Endoh K 1984 A high performance cycloconverter-fed synchronous machine drive system. *IEEE Trans. Ind. Appl.* 20: 1278–1284

Nash J N 1997 Direct torque control, induction motor vector control without an encoder. *IEEE Trans. Ind. Appl.* 33: 333–341

Nilsen R, Kazmeirkowski M P 1989 Reduced order observer with parameter adaptation for first order motor flux estimation in induction machine. *IEE Proc.* D136: 35–43

Noguchi T, Kondo S, Takahasi I 1997 Field oriented control in an induction motor with robust on line tuning of its parameters. *IEEE Trans. Ind. Appl.* 33: 35–42

Nordin K B, Novotny D W, Zinger D S 1985 The influence of motor parameter deviations in feedforward field orientation drive systems. *IEEE Trans. Ind. Appl.* 21: 1009–1015

Novotny D W, Lipo T A 1985 Principles of vector control and field orientation. *IEEE Tutorial Course, Ind. Appl. Soc. Annu. Meet* (Toronto)

Novotny D W, Jansen P L 1991 Field oriented control of synchronous machines. *J. IETE* (India) 37: 46–56

Ogaswara S, Akagi H, Nabae A 1988 The generalised theory of indirect vector control of *ac* machines. *IEEE Trans. Ind. Appl.* 24: 470–478

Okuyama T, Fujimoto N, Fuji H 1990 A simplified vector control system without speed and voltage sensors-effect of setting errors of control parameters and their compensation. *Elec. Eng. Japan* 110: 129–138

Ohm D Y 1989 Simulation of a vector controlled induction motor includes magnetic saturation effects. *J. Intell. Motion PC/M* : 64–79

Ohnishi K, Matai N, Hori Y 1994 Estimation, identification and sensorless control in motion control system. *Proc. IEEE* 82: 1253–1265

Ohnishi K, Uede Y, Miyachi K 1986 Model reference adaptive system against rotor resistance variation in induction motor drive. *IEEE Trans. Ind. Electron.* 33: 217–223

Pallmann R P 1992 First use of a cycloconverter-fed *ac* motor in an aluminium hot strip mill. *Siemens: Energy Autom.* 14: 26–29

Qian J, Rahaman M A 1993 Analysis of field oriented control for permanent hysteresis synchronous motors. *IEEE Trans. Ind. Appl.* 29: 1156–1163

Saito K, Kamiyama K, Sukegawa T, Matsui T, Okuyama T 1987 A multiprocessor-based fully digital *ac* drive system for rolling mills. *IEEE Trans. Ind. Appl.* 23 : 538–544

Sathiakumar S, Biswas S K, Vithyathil J 1986 Microprocessor based field oriented control of a CSI fed induction motor drive. *IEEE Trans. Ind. Electron.* 33: 39–43

Schierling H, Schonherr A 1995 Tough motors under best control: Vector control with SIMOVERT master drives. *Siemens: Drives Control* 3: 17–19

Simoes M G, Bose B K 1995 Neural network based estimation of feedback signals for a vector controlled induction motor drive. *IEEE Trans. Ind. Appl.* 31: 620–629

Sousa G C D, Bose B K 1993 Fuzzy logic based on-line efficiency optimisation control of an indirect vector controlled induction motor. *IEEE /IECON Conf. Proc.* 1168–1174

Stemmler H 1970 Drive system and electronic control equipment of the gearless tube mill. *Brown Boveri Rev.* 57: 120–128

Sugi K, Naito Y, Kurosowa R, Kano Y, Katyama S, Yoshida T 1983 A microprocessor – based high capacity cycloconverter drive for a main rolling mill. *Conf. Proc. Int. Power Electron. Conf.* (Tokyo) 2: 744–755

Tadakuma S, Tanaka S, Naitoh H, Shimana K 1997 Improvement of robustness of vector controlled induction motors using feedforward and feedback control. *IEEE Trans. Power Electron.* 12: 221–227

Tajima H, Matsumoto Y, Umida H, Kawano M 1995 Speed sensorless vector control method for industrial drive system. *Conf. Proc. IPEC* (Yokohama) : 1034–1039

Tanaka H, Nagatani Y, Ehara M 1983 Driving system incorporating vector control inverter for large scale paper machine. *IEEE Trans. Ind. Appl.* 19: 450–455

Terens L, Bourneli J, Peters K 1982 The cycloconverter fed synchronous motor. *Brown Boveri Rev.* 4/5: 122–132

Thakur A N 1996 *On the design, simulation, hybrid implementation and performance assessment of scalar and vector controllers for a PWM inverter fed induction motor drive.* Ph D thesis, Indian Institute of Technology, Kharagpur

Thakur A N, Das S P, De N K, Chattopadhyay A K 1993 Hybrid implementation of indirect vector controlled induction motor and comparison with slip regulated constant V/f control. *Conf. Proc. Natl. Syst. Conf.-93* (Kanpur), pp 261–265

Tiitinen P, Surendra M 1996 The next generation motor control method, DTC direct torque control. *Proc. Int. Conf. Power Electronics, Drives and Energy Systems for Industrial Growth, PEDES'96* (New Delhi) 1: 37–43

Timpe W 1982 Cycloconverter drive for rolling mills. *IEEE Trans. Ind. Appl.* 18: 401–404

Tzou Y Y, Wu Y C 1990 Multimicroprocessor based robust control of an *ac* induction servo motor. *IEEE Trans. Ind. Appl.* 26: 441–449

Vas P 1990 *Vector control of ac machines* (New York: Oxford University Press)

Vas P, Alakula M 1990 Field oriented control of saturated induction machine. *IEEE Trans. Energy Conversion* 5: 218–224

Verghese G C, Sanders S R 1988 Observers for flux estimation induction machines. *IEEE Trans. Ind. Electron.* 35: 85–94

Wescheta A 1983 Design considerations and performances of brushless permanent magnet servo motors. *IEEE Ind. Appl. Soc. Annu. Meet Conf. Rec.*, pp 469–475

Wu Z K, Strangas E G 1988 Feed forward field orientation control of an induction motor using a PWM voltage source inverter and standard single board computer. *IEEE Trans. Ind. Electron.* 35: 75–79

Xu L, Cheng W 1995 Torque and reactive power control of a doubly fed induction machine by position sensorless scheme. *IEEE Trans. Ind. Appl.* 31: 636–642

Xu X, Novotny D W 1991 Implementation of direct stator flux orientation control on a versatile DSP based system. *IEEE Trans. Ind. Appl.* 27: 694–700

Xu X, Novotny D W 1992 Selection of flux reference for induction machines in the field weakening region. *IEEE Trans. Ind. Appl.* 28: 1353–1358

Xu L, Xu X, Lipo T A, Novotny D W 1991 Vector control of a synchronous reluctance motor including saturation and iron loss. *IEEE Trans. Ind. Appl.* 27: 977–984

Yanagawa K, Sakai K, Ishida S, Endou T, Fujii H 1995 Autotuning general purpose inverter with sensorless vector control. *Conf. Rec. Int. Power Electron. Conf. 95* (Yokohama), pp 1005–1009

Zinger D S, Profumo F, Lipo T A, Novotny D W 1990 A direct flux orientation controller for induction motor drives using tapped stator windings. *IEEE Trans. Power Electron.* 5: 446–453

Zhao Y, Lipo T A 1996 Modelling and control of a multiphase induction machine with structural unbalance. Part-II. Field oriented control and experimental verification. *IEEE Trans. Energy Conversion* 11: 578–584

Switched reluctance motor drives – recent advances

M EHSANI

Texas A& M University, Department of Electrical Engineering, College Station, TX 77843, USA
e-mail: ehsani@ee.tamu.edu

Abstract. The objective of this paper is to review the state-of-the-art and recent developments in Switched Reluctance Motor (SRM) drives. The interest for improved performance and reliability has motivated many SRM advances in the recent years. Even after almost 30 years of research in SRM, which might appear to be the simplest of all machines, there remain critical issues to be explored to gain deeper insight into the SRM technology. The paper briefly discusses the historical background and the basic operating principles of the motor. The topics discussed include the current state of research in converter topologies, control algorithms, torque ripple, noise, and sensorless operation. Recent advances in the field of SRMs indicates that they will have an increasing influence in the area of variable speed drives in the coming decades.

Keywords. Switched reluctance motors; converter topologies; control algorithms.

1. Introduction

Switched Reluctance Motor drives (SRMs) are relatively new entrants in the rapidly developing variable speed drives market. They are inherently variable speed drives which have simple construction, wide speed ranges, good energy efficiencies, high torque to inertia ratios, and high torque to power density ratios. The simple structure of SRMs will likely make them less expensive than the equivalent variable speed drives in mass production. An SRM has the flexibility of operating as a four-quadrant drive with independent control of speed and torque over a wide speed range. Their wide torque and speed range eliminates the need for expensive and troublesome mechanical gears and transmissions. Figure 1 shows a selection of SRMs with power ranging from 100 W to 75 kW and speeds ranging of 250 to 30000 rpm (Lawrenson 1992).

1.1 *Historical background*

SRM is the modern version of the 'electromagnetic engine' which dates back to the late 1830's. The modern era of developmental SRM started in 1972 when it was patented by
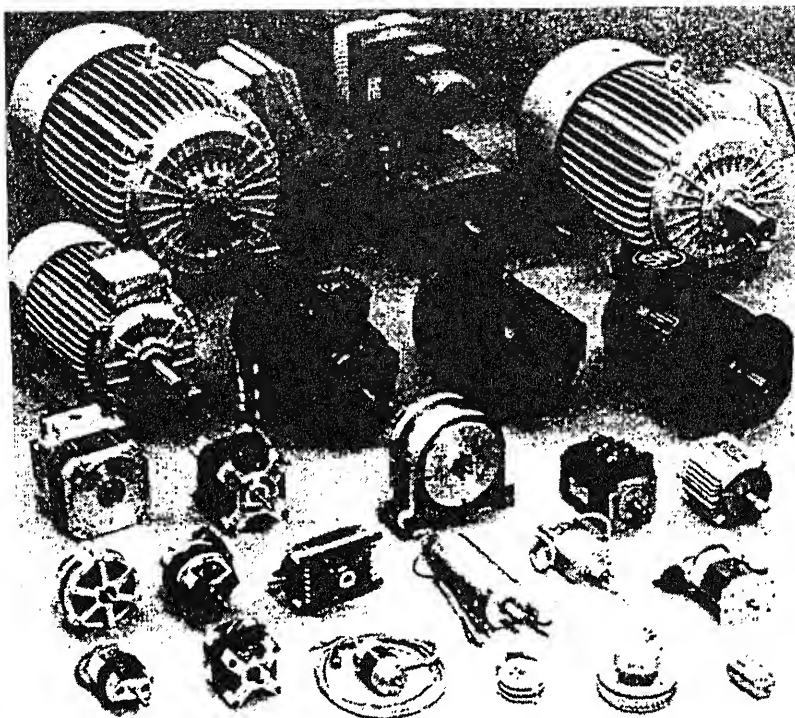
Figure 1. Assortment of SRMs.

Bedford. SRMs received considerable attention after the exemplary work done at the Universities of Leeds and Nottingham in the 1980s (Ray & Davis 1979; Lawrenson *et al* 1980; Davis *et al* 1981). This spurred a series of research efforts all over the world, especially in Europe and in the US, resulting in several publications, patents and applications (Lawrenson 1992) which is illustrated in figure 2. Even after almost thirty years of research in SRM, which appears to be the simplest of all machines, there remain some critical issues which need further study.

1.2 *Basic principles of operation*

SRMs are structurally similar to Variable Reluctance Stepper motors, but they differ in the following aspects:

- stator phase currents are switched based on the rotor position feedback;

- the machine is designed to operate efficiently for a wide range of speed.

SRM has salient poles on both stator and rotor with concentrated windings only on the stator and no windings on the rotor. Windings on the diametrically opposite stator poles are connected in series. Figure 3 shows a typical 8/6 SRM. Currents in the stator windings are switched on and off in accordance with the rotor position feedback.

The basic principle of operation of SRM is like any other reluctance motor – torque is produced by the tendency of the rotor to align itself to the minimum reluctance position. As this is independent of the direction of the current in the stator windings, the power converter
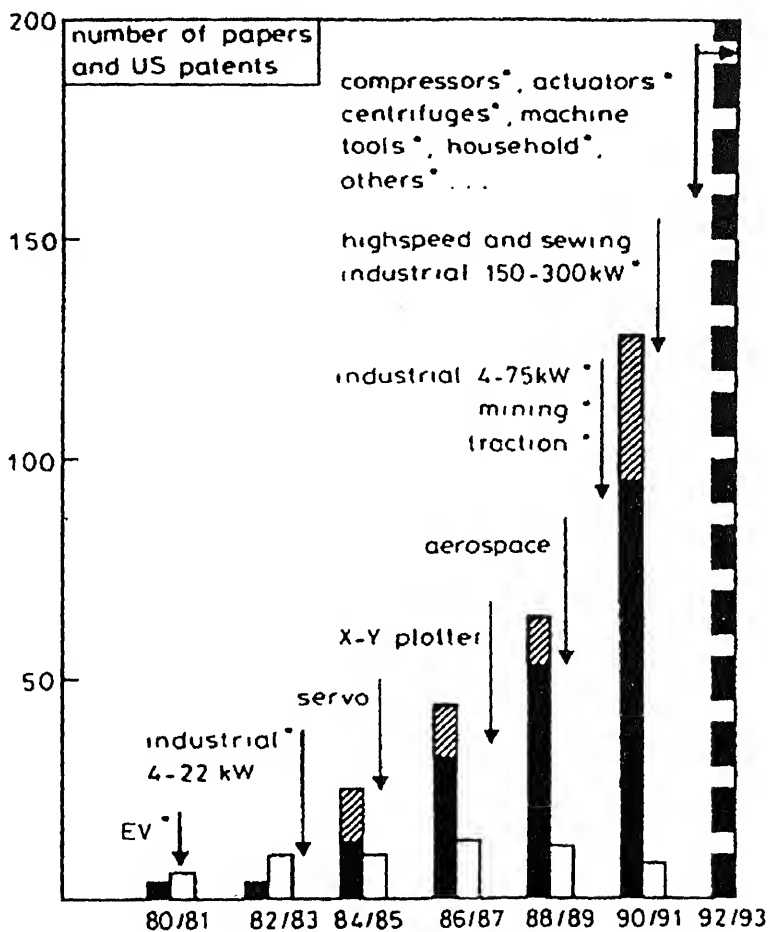
Figure 2. Approximate number of publications, patents and applications.
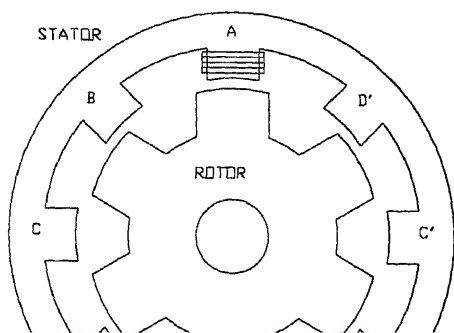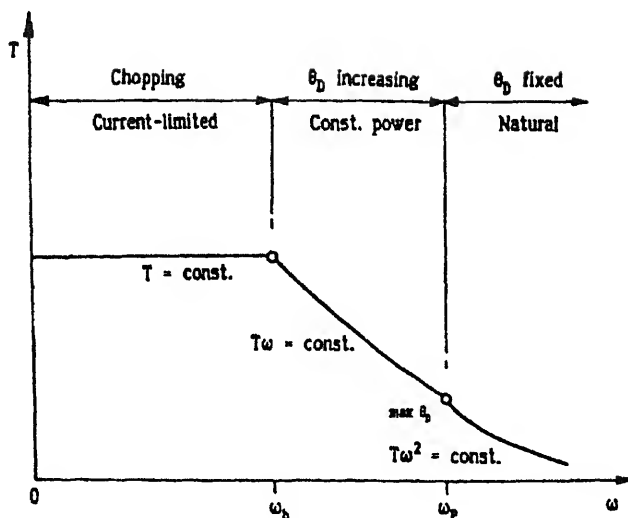
Figure 4. Torque-speed characteristics of an SRM.

circuit can be simplified. The radial magnetic attraction that operates the SRM is about ten times larger than the circumferential forces that operate an induction motor. The torque-speed characteristics of a SRM replicates a conventional DC machine characteristics under fixed firing switching strategy. Typical torque speed characteristics and variation in the flux linkage with current at different rotor positions are shown in figures 4 and 5.

2. Advances in converter design

SRMs cannot operate directly with a *dc* supply or the standard sinusoidal *ac* supply available 'off the wall'. Hence it is important that the design of the converter should be
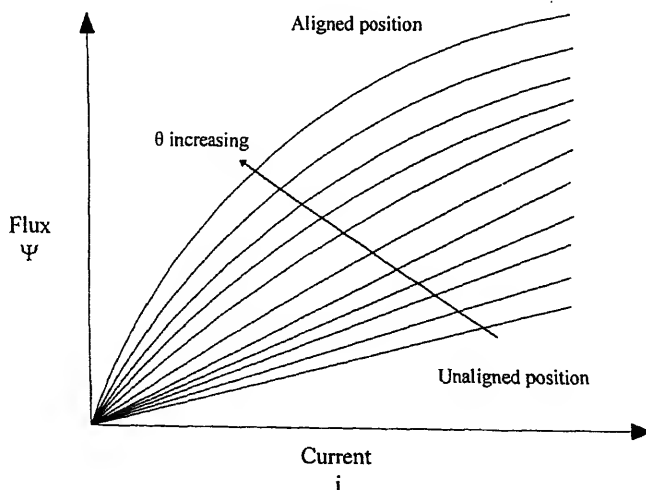
Figure 5. Flux/current/position curves.

coordinated concurrently with the design of the motor to obtain an optimal design of the drive as a whole. Unlike the motors that operate with sinusoidal voltages and currents, the converter topology in an SRM is dependent on the machine design. The topology depends on the motor configuration, the number of stator and rotor poles and the associated conduction and overlap angle (Vukosavic & Stefanovic 1991). In general, the starting torque requirements decides the current rating and the maximum speed of the motor decides the voltage rating of the inverter.

An ideal converter must satisfy

- low switches per phase ratio,
- ability to supply and control a commanded current independently and precisely,
- flexibility in adapting to any number of phases (odd or even),
- low VA rating for a given rating of the drive,
- robustness and reliability,
- good efficiency,
- ability to operate in all four quadrants effectively,
- less torque ripple and noise.

The most commonly used converter types are the classic half-bridge type converters and the split phase type converters. The classic type converter is the most flexible type converter but it requires a greater number of switches. The split phase type converter requires an even number of phases and has high active device rating, and hence is suitable only for low voltage and low power applications. Another type of converter suitable for star connected type SRM is the dual-decay type converter with flexibility of control in freewheeling mode and reduced device ratings (Ehsani *et al* 1993). Most of the converters developed recently are aimed at reducing the number of switches and are found to be more application-specific with some trade-offs. The different converter topologies are shown in figure 6.

3. Advanced control schemes

A reliable SRM drive system must have the following characteristics
- parameter insensitive control characteristics
- quick precise dynamic response with no overshoot
- rapid recovery from transient disturbances

Conventional linear controllers are quite sensitive to plant parameters. This along with the need for optimum performance of SRM like maximum efficiency, maximum torque and minimum torque ripple calls for sophisticated control strategy. The important control variables in an SRM, switch-on and switch-off angles, are a nonlinear complex function of many motor parameters and thereby require fine tuning (sophisticated control) for optimal
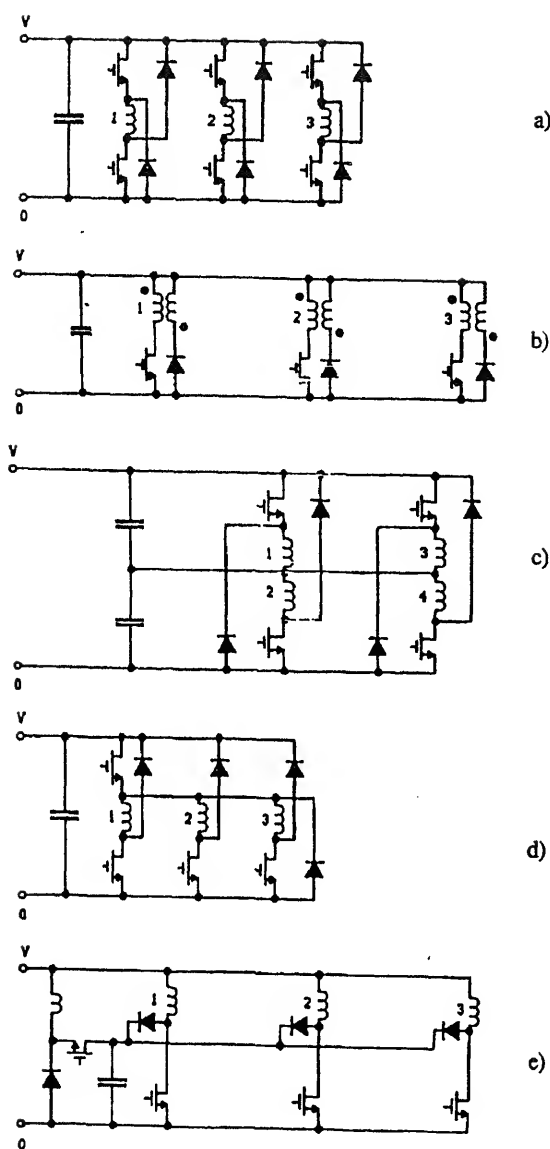
Figure 6. SRM converter topologies. (a) Bridge, (b) bifilar, (c) split capacitor, (d) Miller, and (e) C-dump circuits.

surface in the state space. These methods are shown to provide a better torque ripple characteristics and are insensitive to parameter variations and disturbances. Artificial neural networks (ANNs) have been used successfully in the control of nonlinear dynamic systems (Reay *et al* 1993). The capability to accommodate accurate nonlinear modeling has made Artificial Neural Networks (ANNs) an ideal candidate to solve the control strategies of inherently nonlinear SRMs. Fuzzy logic controllers are gaining interest recently in the field of nonlinear control (Bolognani & Zigliatto 1993). They offer the following advantages — they do not require an accurate model of the plant, they can be designed on the basis of linguistic information obtained from the previous knowledge of the control of the machine and give better performance results than the conventional controllers. Amor *et al* (1992)

suggested an adaptive control based on feedback linearization for less torque ripple (only position control) later used fifth order model (Amor *et al* 1993) to control speed and torque for adaptive control. A machine independent method control to minimize energy consumption is discussed by Kjaer *et al* (1994).

3.1 *Need for self-tuning control*

The following are the assumptions (Ehsani & Ramani 1993) made in most of the conventional control methods:

- the inductance is symmetrical about the aligned position,
- all the rotor and stator poles have perfect symmetrical tooth structures,
- the inductance variation is time and temperature independent,
- saturation has no or very less effect on the inductance variation,
- the inductance profile is identical for all the phases,
- the inductance profile is identical for all the machines of the same rating in mass production.

In practice the above assumptions are not valid and these factors affect the sensitivity of a control algorithm. Control methods that can adapt to the parameter drifts are required to optimize the machine performance. Recently the problem of obtaining optimal performance from an SRM in the presence of parameter variation has gained considerable interest (Tandon *et al* 1996). New self-tuning algorithms which optimize the steady-state performance of the drive as measured by Torque per Ampere (TPA) have been introduced. It has been realized only recently that due to parameter variation and drift, the phase inductance profiles can significantly differ from the design data (Ehsani & Ramani 1993; Ehsani *et al* 1993). Hence, it becomes necessary to use a controller with self-tuning capability if optimal performance of the SRM drive is to be maintained. Maximum TPA is desirable for any drive application because the motor may be described as a 'current to torque transducer'. Specifically, the following problems arise in practical SRMs.

- Due to manufacturing tolerances the inductance profile varies by as much as 10% from phase to phase of the motor and also from motor to motor with the same design and rating. It is noted that the minimum inductance does not show any significant variation. This is due to the very large air-gap at the unaligned position. However, the maximum inductance occurs at the aligned position where the air gap may be less than 1mm. Therefore any minor variations in the air-gap show up as change in the maximum inductance for each phase. Thus, the slope of the inductance profiles will also differ and the torque production is affected.

- With time, there will be wear on the bearings, as a result of which the air-gap may change or acquire a small eccentricity. Again, this has an impact on the maximum inductance of each phase.
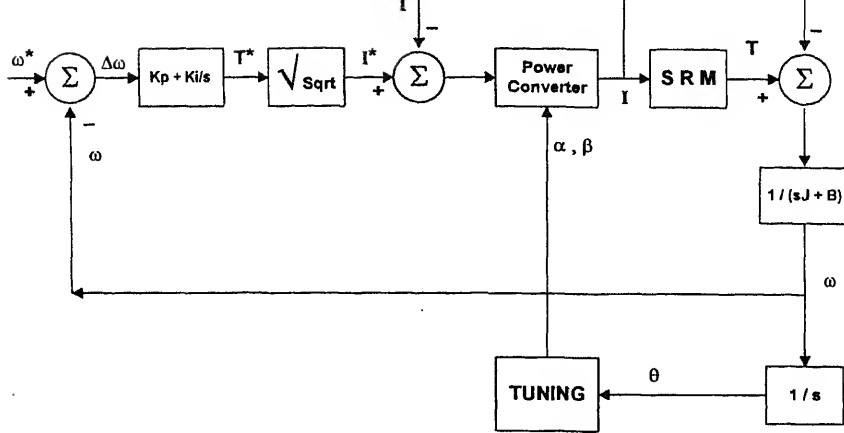
Figure 7. Control block diagram with self-tuning algorithm.

Hence, the optimal values of the turn-on angle calculated off-line are sufficient for the TPA maximization. An on-line, self-tuning algorithm to determine the optimal value of the turn-off angle in the presence of parameter variations which alter the inductance profiles has proven to produce superior steady state performance. Figure 7 shows the block diagram of the above mentioned self-tuning control. The inherent simplicity of this new approach makes it ideal for real-time implementation in a digital control system. The control scheme is applicable to any SRM drive operated with a shaft position sensor, and does not depend on the number of phases, poles or HP of the motor.

4. Advances in sensorless operation

Closed loop operation of an SRM requires rotor position information for satisfactory performance. Conventional methods used for position sensing include resolvers, inductive or Hall effect sensors and optical encoders. These have the disadvantages like additional cost, additional electrical connections, mechanical alignment problem, less suitability to space restricted application and the significant disadvantage of being a potential source of unreliability. These lead to the research in sensorless operation of SRMs, resulting in several techniques in the past two decades. Most of the existing methods extract the rotor position information from the measurable electrical parameters. These techniques eliminate the requirement of the conventional position sensors, thereby increasing the reliability of the motor drive system considerably. They can be classified as below.

(1) *Active probing methods:* These methods utilize the responses of diagnostic signals injected into the passive (un-energized) phase of an SRM. These methods are suitable for low speeds as the time window for passive phase measurement reduces at high speeds. Usually, the phase to be energized next is diagnosed for position estimation. They can be further classified as (a) Linear relation methods, and (b) inverse relation methods

(a) Linear relation methods – In these methods, the signal containing position information is directly proportional to the phase inductance. In a typical method the rate of change of phase current, which is influenced by the incremental inductance, is monitored (Acarnley *et al* 1985). Rotor position can now be deduced as the incremental inductance is a function of rotor position. This has an advantage of deducing the rotor position even at zero speed. Another method which is robust to switching noise was presented by Ehsani and coworkers (Ehsani & Ramani 1994; Ehsani *et al* 1994) and called the PM encoder technique. A sinusoidal carrier voltage signal of frequency much higher than that of the frequency of variation of inductance is chosen. Thus, the transient variation of the current phase will contain the information about the dynamic motor winding inductance. This encoded inductance information is decoded using zero crossing detectors for the voltage and current. The demodulator generates a square wave signal whose pulse width variation represents the phase inductance variation.

Also a modified PM encoder technique is presented which is suitable for a wide range of speeds. Mathematical analysis and simulation results show that PM technique is more sensitive for lower values of inductance and AM technique is more sensitive for higher inductance values. To achieve a better sensitivity, a level crossing detector is used instead of a zero crossing detector. The level crossing detector is set to a threshold value. Now the square wave output corresponds to the phase angle variation with respect to the threshold value other than zero which gives better sensitivity than the above PM method.

(b) Inverse relation methods – In these methods, the position information encoded signals are inversely proportional to the phase inductance. At high speeds, the motional EMF is very high, the current will never reach the rated value resulting in a 'single pulse' mode. In this mode, the current gradient in the next phase to be excited has the position information (Acarnley *et al* 1985). At the turn-on position, as the resistive voltage drop and the motional EMF are negligible, the initial rate of change of current is inversely proportional to the incremental inductance which gives the position information directly. The sensorless operation is implemented by comparing the initial current gradient with a optimal current gradient. In another method called the Amplitude Modulation (AM) method (Ehsani *et al* 1990; Ehsani 1991), the position information can be obtained from the amplitude of the current as it is directly proportional to the inductance variation. In this method, the envelope of the modulated current signal is detected. In addition, the information can be decoded by measuring the amplitude in terms of angles using a level crossing detector.

(2) *Non-intrusive methods:* In these methods, the rotor position is obtained based on the measurable parameters without using any diagnostic or probing signals. Neglecting R, at low speeds the incremental inductance $l = d\psi/di$ is a function of θ for constant i. Therefore, 'dt', rise time or fall time, can be used to obtain the position information. The flux linkage curve information on a multi-dimensional look-up table can be used to determine rotor position (Hedland 1986, 1992). In a more recent method called active phase vector method (Ehsani *et al* 1994), a composite vector which is directly proportional to the inductance is obtained based on the discrete form of voltage equation for different

(3) *Open loop or synchronous control method:* These methods are based on synchronous control and do not actually provide any position information. The motor is run from a variable frequency oscillator and change is made only to the dwell angle to improve the stability (Miller *et al* 1985).

(4) *Other methods:* Some other methods are as below.

(a) *Observer based* – Observer method reconstructs the state of the SRM drive system on the basis of known system inputs and system measurements (Lumsdaine & Lang 1990). Measurements of input voltages and currents were utilized. An accurate mathematical model including mechanical load (in state space form) to estimate current, flux linkage, speed and rotor position which was compared with actual current and error adjustment made using an adjustment matrix to estimate the position is used.

(b) *Mutually induced voltage based methods* – In this method, the mutually induced voltage in an un-energized phase due to current in energized phase is monitored to obtain the position information (Hussain & Ehsani 1994).

(c) *Design based method* – This method is based on altering slightly the structure of at least one of stator and/or rotor pole faces which will introduce a perturbation in the inductance profile of the motor while it is running (Bartos *et al* 1993). The perturbations can be produced by introducing a notch or bump in the stator and/or rotor pole faces. The frequency of these perturbations gives direct information on the speed of the motor.

The corruption of position information due to the secondary effects of the existing sensorless methods must be considered to improve accuracy. It is found that a trade-off exists between extensive computation and good resolution in sensorless position sensing. There is still much room for improvement as there are no accurate, commercially applicable methods available. The need for inexpensive, reliable, indirect position sensing technique for a wide range of speed, still exists. The advances made in the field of power electronics, motion control and signal processing can be used to improve the commercial applicability of the existing methods. The existing computationally intensive, high resolution sensorless techniques can be made commercially viable in the future with the advances made in the computational power of Digital Signal Processors. Further development resulting in commercially applicable, inexpensive techniques is expected. More research will be necessary for a method with good positional accuracy and suitable for commercial application.

5. Overview of critical issues

Apart from the numerous advantages, SRMs are also known for their high torque pulsations, high acoustic noise and reliability issues due to sensor based operation.

5.1 *Torque ripple*

The nonlinear coupling between the rotor position, phase current and overlap angle and the doubly salient geometric structure of the SRM are the intrinsic causes of torque ripple in a SRM. Torque ripple is very undesirable in low speed and servo type applications. Several methods have been developed to reduce torque ripple based on machine design or control strategies.

5.1a *Design based methods:* Torque ripple can be minimized by suitably designing the magnetic structure of the machine (Tormey & Torrey 1991). Comprehensive procedures, beginning with the fundamental selection of pole numbers and geometry, for designing SRM drives for low torque ripple applications has been presented in the literature (Wallace & Taylor 1990). Values for specific torque are used to estimate the required SRM size and are obtained from empirical data or from an analytical estimation method. Pole numbers are chosen based on the speed and torque-ripple specifications for the design. The pole numbers define a range of feasible pole arc combinations. The centre pole arc values are chosen as a starting point. The current density in the phase winding is also chosen based on the thermal constraint of the application. The pole arcs and motor dimensions which yield minimum ripple are selected as the candidate design which is further evaluated using the dynamic SRM model.

5.1b *Control based methods:* Classical linear controllers cannot eliminate the torque pulsations. The fundamental approach is by optimal current profiling that reduces torque pulsations. Torque ripple can be reduced by using a current-tracking control method in which the desired stator currents are computed by linearizing and decoupling transformation. The shape of static torque-angle-current characteristics of SRM drive can be fully determined by a series of measurements performed with the drive in a self-learning mode (Kavanagh *et al* 1991). Based on this, the current required to obtain the optimum torque contribution from each phase, at each rotor position, can be determined for a smooth torque performance. A single input, linear, decoupled output torque controller based on optimal precalculation of the phase current profile provides low torque ripple (Schramm *et al* 1992). A bi-cubic spline interpolation was used to model the nonlinear experimental data. This method optimizes the current overlap at all torque levels so as to minimize the peak phase current. This current profiling algorithm results in the highest possible operating speed range under constant torque operation. The torque output is decoupled single-input linear function of torque input demand. Neural techniques can learn the current profiles required to minimize torque ripple and to satisfy other performance criteria on-line (Reay *et al* 1993). Torque measurement is required to train the neural network. PWM current control can be used for smooth operation of an SRM drive (Hussain & Ehsani 1994). The torque pulsations during commutation are minimized by a current control strategy which allows simultaneous conduction of two positive torque producing phases over an extended predefined region. The effects of saturation can also be taken into account.

Acoustic noise levels in SRMs are relatively high when compared to other *ac* drives. The acoustic noise has both magnetic and mechanical origin (Cameron 1992) . The possible sources include radial attractive forces between the rotor and stator, stator vibrations induced by the torque ripple, stator winding vibrations induced by the interaction of the stator current and the local magnetic field, magnetostrictive forces in the stator laminations, unbalanced magnetic and mechanical forces on the rotor due to manufacturing asymmetry windage and bearing vibrations. Of these, the dominant one is shown to be the radial vibrations of the stator. Solving the noise problem can be approached either from the motor design point of view or the control point of view or a combination of these two. Acoustic noise can be reduced to any required level by current shaping or by introducing dither into the turn-on and turn-off angles. In another technique a chopper is introduced between rectifier and converter to reduce the phase voltage with respect to the speed (Pillay *et al* 1994). The voltage smoothing method reduces the rate of change of radial force and produces a smaller vibration (Pollock & Wu 1995). A three-stage commutation technique has been described in the literature which cancels the stator vibrations when the power converter does not have a zero volt loop by employing the three stage commutation technique at the beginning and end of the zero volt loop. The active cancellation methods like the three stage commutation technique are superior to the voltage smoothing method since they allow the energy to be dissipated in subsequent vibrations which can completely oppose each other. Research in this area is still in its infancy and further developments are expected in the near future.

6.   Current state of research

6.1   *SR generators*

Fukao (1986) discussed the principles and output characteristics of a super high speed reluctance generator system. Cameron *et al* (1992) discussed the computer-aided design of a VR generator and later about the control aspects of a high speed VR generator (Cameron & Lang 1992) system. Radun (1994) discussed the analysis and design of an aircraft engine starter/generator. He proposed different excitation systems including one for three phase generation. Torrey & Hassanin (1995) present the design methodology for low-speed VRGs for renewable energy systems.

6.2   *Fault tolerance*

SRMs inherently have a good fault tolerance capability in the sense that they can operate with faulted motor windings or inverter circuitry. The magnetic independence of the motor phases and the circuit independence of the SRM converter gives SRM that capability. Stephens (1991) discussed the different types of faults that can occur and suggested a method to sense and isolate the faulted winding. Miller (1995) analysed SRM under faulted condition using FEM, magnetic modeling, actual experiments as well as a simulation software.

6.3 *Losses and temperature rise*

Core losses in SRM forms a significant part of the total losses. Materu & Krishnan (1988) presented a method of calculating the iron losses based on the flux density waveform including saturation and presented a computationally intensive procedure using FEM time-stepping. Metawally *et al* (1988) discussed the core loss in SRM based on experimental results. Faiz & Dadgari (1991) analysed the heat distribution and thermal calculations for multitooth SRM under natural and forced cooling condition and compared it with experimental results. Corda & Olaca (1993) predicted (analysed) losses in the converter of SR drive under different modes of operation. The prediction of converter losses, for a given converter topology and method of control, is useful when selecting the type of power semiconductor switching components and design parameters of control with respect to efficiency of the drive. The prediction of converter losses is also useful for the correct thermal design. A new approach to calculate core losses was developed by Hayashi & Miller (1994) in matrix form using the Steinmetz equation.

7. Applications

SRMs have been successfully applied to a variety of applications resulting in high performance drives. Figure 8 shows typical efficiency vs. speed for curves for various 30 HP motors. The first successful application was as general purpose industrial drives. More recently, they have been commercially applied to automotive utility motors. Good fault tolerance and ability to operate in harsh environments made SRMs successful candidates as coal-shearing machines, textile spinning drives, friction welding machines, food processor applications, plotters, aerospace and automotive applications. SRMs of 5 MW at
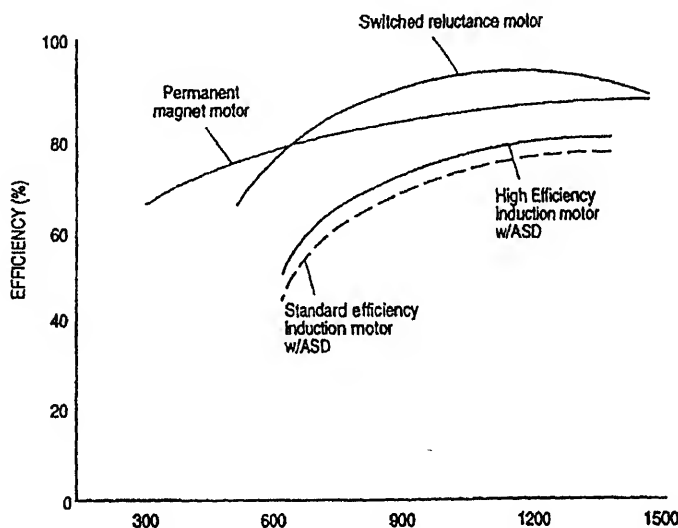
50 rpm to 10 kW at 100,000 rpm have been built and tested successfully. Speed range of a SRM can be from 100 : 1 to 1000 : 1.

The fault-tolerance capability of SRMs is extremely good which makes it suitable for aerospace, automotive and industrial applications. The independence of each phase windings and the absence of shoot-through paths contributes to the fault tolerance of the SRMs. Typical applications include traction, domestic appliances, mining, servo type and battery-powered applications. The application areas of SRM are rapidly expanding as the SRMs can compete virtually in any industrial or domestic drive market.

8.   Conclusions

The state-of-the-art SRM drive technology and the recent advances made in the research of SRM were discussed. A wide range of topics including converter topologies, control algorithms, sensorless operation, fault tolerance, noise and vibration issues were presented. From the above discussions it can be clearly seen that the SRM drive technology has come a long way. Further research is necessary in some critical areas for the SRM to be made commercially acceptable as a variable speed drive. Recent advances in power electronics and control systems are set to make SRM a commercially acceptable drive in the immediate future.

References

Acarnley P P, Hill R J, Hooper C W 1985 Detection of rotor position in stepping and switched motors by monitoring of current waveforms. *IEEE Trans. Ind. Electron.* IE-32: 215–222

Amor L B, Dessaint L A, Akhrif O, Olivier G 1992 Adaptive feedback linearization for position control of a switched reluctance motor: Analysis and simulation. *IEEE Ind. Electron. Conf.* pp 150–159

Amor L B, Dessaint L A, Akhrif O, Olivia G 1993 Adaptive input-output linearization of a switched reluctance motor for torque control. *IEEE Ind. Electron. Conf.*, pp 2155–2160

Bartos, Houle T H, Johnson J H 1993 Switched reluctance motor with sensorless position detection. US Patent No. 5,256,923

Bolognani S, Zigliotto M 1993 Fuzzy logic control of a switched reluctance motor drive. *IEEE Ind. Appl. Spec. Conf. Rec.*, pp 2049–2054

Cameron D E 1992 The origin and reduction of acoustic noise in doubly salient variable-reluctance motors. *IEEE Trans. Ind. Appl.* 28: 1250–1255

Cameron D E, Lang J H 1992 The control of high-speed variable-reluctance generators in electric power systems. *IEEE Power Electron. Spec. Conf.*, pp 121–125

Cameron D E, Lang J H, Belanger D 1992 The computer-aided design of variable-reluctance generators. *IEEE Appl. Power Electron. Conf.*, pp 114–120

Corda J, Olaca M 1993 Analysis of losses in power electronic converter of SR drive. *Fifth Eur. Conf. on Power Electronics and Applications*, pp 49–53

Davis R M, Ray W F, Blake R J 1981 Inverter drive for switched reluctance: circuits and component ratings. *Inst. Elec. Eng. Proc.* B128: 126–136

Ehsani M 1991 Position sensor elimination technique for the switched reluctance motor drive. US Patent No. 5,072,166

Ehsani M, Ramani K R 1993 Direct control strategies based on sensing inductance in switched reluctance motors. *IEEE Power Electron. Spec. Conf. Rec.*, pp 10–16

Ehsani M, Ramani K R 1994 New commutation methods in switched reluctance motors based on active phase vectors. *IEEE Power Electron. Spec. Conf. Rec.*, pp 493–499

Ehsani M, Hussain I, Kulkarni A B 1990 Elimination of discrete position sensor and current sensor in switched reluctance motor drives. *IEEE Ind. Appl. Spec. Conf. Proc.* 518–524

Ehsani M, Husain I, Ramani K R, Galloway J H 1993 Dual-decay converter for switched reluctance motor drives in low-voltage applications. *IEEE Trans. Power Electron.* 8: 224–230

Ehsani M, Hussain I, Mahajan S, Ramani K R 1994 New modulation encoding techniques for indirect rotor position sensing in switched reluctance motors. *IEEE Trans. Ind. Appl.* 30: 85–91

Faiz J, Dadgari A 1991 Heat distribution and thermal calculation for switched reluctance motors. *Fifth Int. Conf. Electrical Machines and Drives, Conf. Publ.* No. 341, pp 305–310

Fukao T 1986 Principles and output characteristics of super high-speed reluctance generation system. *IEEE Trans. Ind. Appl.* 22: 702–707

Hayashi Y, Miller T J E 1994 A new approach to calculating core losses in the SRM. *IEEE Ind. Appl. Spec. Conf. Rec.*, pp 322–328

Hedland M 1986 A method and a device for sensorless control of a reluctance motor. *Int. Patent* No. WO 91/02401

Hedland M 1992 Method and a device for sensorless control of a reluctance motor. US Patent No. 5,173,650

Hussain I, Ehsani M 1994 Torque ripple minimization in switched reluctance motor drives by PWM current control. *IEEE Appl. Power Electron. Conf.*, pp 72–77

Hussain I, Ehsani M 1994 Rotor position sensing in switched reluctance motor drives by measuring mutually induced voltages. *IEEE Trans. Ind. Appl.* 30: 665–672

Kavanagh R C, Murphy J M D, Egan M G 1991 Torque ripple minimization in switched reluctance drives using self-learning techniques. *IEEE Ind. Electron. Conf.*, pp 289–294

Kjaer P C, Nielson P, Anderson L, Blaabjerg F 1994 A new energy optimizing control strategy for switched reluctance motors. *IEEE Appl. Power Electron. Conf.*, pp 48–55

Lawrenson P J 1992 Switched reluctance drives: A perspective. *Proc. Int. Conf. Elec. Machines* 1: 12–21

Lawrenson P J, Stephenson J M, Blenkinsop P T, Corda J, Fulton N N 1980 Variable-speed switched reluctance motors. *Inst. Elec. Eng. Proc.* B127: 253–265

Lumsdaine, Lang J H 1990 State observers for variable-reluctance motors. *IEEE Trans. Ind. Electron.* 37: 133–142

Materu P, Krishnan R 1988 Estimation of switched reluctance motor losses. *IEEE Ind. Appl. Spec. Conf. Rec.*, pp 79–89

Metawally H M B, Faiz J, Finch J W 1988 Core loss in switched reluctance motor structure – experimental results. *Proc. Int. Conf. Electrical Machines*, Pisa, Italy, 2: 31–34

Miller T J E 1993 Faults and unbalance forces in the switched reluctance machine. *IEEE. Ind. Appl. Spec. Conf. Rec.*, pp 87–96

Miller T J E, Bass J T, Ehsani M 1985 Stabilization of variable-reluctance motor drives operating without shaft position sensor feedback. *Incremental Motion Control Systems and Devices Proceedings*, pp 361–368

Pillay P, Samudio R, Ahmed M, Patel R 1994 A chopper-controlled SRM drive for reduced acoustic noise and improved ride-through capability using super capacitors. *IEEE Ind. Appl. Spec. Conf. Rec.*, pp 313–321

Pollock C, Wu C Y 1995 Acoustic noise cancellation techniques for switched reluctance drives. *IEEE Ind. Appl. Spec. Conf. Rec.*, pp 448–455

Radun A 1994 Generating with switched reluctance motor. *IEEE Appl. Power Electron. Conf.*, pp 41–47

Ray W F, Davis R M 1979 Inverter drive for doubly-salient reluctance motor: its fundamental behavior, linear analysis and cost implications. *IEEE Elec. Power Appl.* 2: 185–193

Reay D S, Green T C, Williams B W 1993 Neural networks used for torque ripple minimization from a switched reluctance motor. *Fifth European Conf. on Power Electronics and Applications* 1–6

Schramm, Williams B W, Green T C 1992 Torque ripple reduction of switched reluctance motors by phase current optimal profiling. *IEEE Power Electron. Spec. Conf.* 856–860

Stephens C M 1991 Fault detection and management system for fault-tolerant switched reluctance motor drives. *IEEE Trans. Ind. Appl.* 27: 1098–1102

Tandon P, Rajarathnam A V, Ehsani M 1996 Self-tuning control of a switched reluctance motor drive with shaft position sensor. *IEEE Ind. Appl. Spec. Conf. Rec.* 101–108

Tormey D P, Torrey D A 1991 A comprehensive design procedure for low torque-ripple variable reluctance motor drives. *IEEE* 244–251

Torrey D A, Hassanin M 1995 The design of low-speed variable reluctance generators. *IEEE Ind. Appl. Spec. Conf. Rec.* 427–433

Vukosavic S, Stefanovic V R 1991 SRM inverter topologies: a comparative evaluation. *IEEE Trans. Ind. Appl.* 27: 1034–1047

Wallace R S, Taylor D G 1990 Three-phase switched reluctance motor design to reduce torque ripple. *Proc. Int. Conf. Electrical Machines* Cambridge, MA, pp 783–787

Recent advances in permanent magnet brushless DC motors

BHIM SINGH

Department of Electrical Engineering, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India
e-mail: bsingh@ee.iitd.ernet.in

Abstract. This paper deals with the latest developments in Permanent Magnet Brushless DC (PMBLDC) motor drives. A comprehensive account of the state-of-the-art on types of construction of the motor, closed loop controllers in position, speed and current/torque control and recent trends in inverters, sensors etc. are given. Techniques for mechanical sensors elimination are discussed in detail. Special efforts made to reduce torque ripples, noise and vibrations are described. The impact of microelectronics through integrated chips used in the control of PMBLDC motor drives is given. The increasing applications of this drive due to improved performance and its cost reduction are also enlisted.

Keywords. PMBLDC motor; sensors; controllers; sensorless operation; torque pulsations.

1. Introduction

Permanent Magnet Brushless DC (PMBLDC) motors are increasingly being used in a wide spectrum of applications such as domestic equipments, automobiles, information technology equipment, industries, public life appliances, transportation, aerospace, defence equipment, power tools, toys, vision and sound equipment and medical and health care equipment ranging from microwatts to megawatts [1–24]. It has become possible because of their superior performance in terms of high efficiency, fast response, light weight, precise and accurate control, high reliability, maintenance free operation, brushless construction, high power density and reduced size. Recent developments in PMBLDC motor technology in terms of availability of high performance rare earth PM materials, varying motor constructions such as axial field, radial field, package type, rectangular fed, sine fed motors, improved sensor technology, fast semiconductor modules, low cost high performance microelectronics devices, new control philosophy such as robust, adaptive, fuzzy, neural AI based controllers, have been a boon to their widespread use in the large speed ranges from few revolutions to several thousand revolutions per minute (rpm). They have been proven most suitable for position control in machine tools, robotics and high

NB: Citations in this paper are not in our usual format.

precision servos, speed control and torque control in various industry and process control applications [7,8,21-24].

Inspite of being one of the best, the PMBLDC motor has faced many hurdles to come to its present stage in terms of cost, torque ripples, noise, vibrations, reduced reliability due to the large number of components, operational constraints such as temperature rise etc. Continuous efforts have been made to overcome these problems on different aspects of this drive. The PMBLDC motor drive is undoubtedly quite a big mission in itself; this paper concentrates on the recent advances in PMBLDC motors in terms of motor construction, closed loop controllers, semiconductor power modules, sensors and their reduction, torque ripple minimization, impact of microelectronics, cost reduction and potential applications.

2. Latest developments in the PMBLDC motors

Permanent magnet (PM) excitation has been used in place of *dc* excitation in different electric machines such as *dc* machines, synchronous machines and new PMBL machines such as PM stepper motors, hybrid stepper motors and PMBLDC motors. High cost of PM materials has been a major bottleneck for use and development of these electric machines. Gradual growth of better PM materials, improved manufacturing technology, varying nature of construction of these motors to suit specific applications have brought them at a level where they are considered one of the best motors available nowadays. PM machines have a wide spectrum but this paper is restricted to PMBLDC motors.

Presently PM materials used in PMBLDC motors are classified in the following three broad categories [8-10], namely Alnico (Al-Ni-Co-Fe), Ceramics also include ferrites and rare-earth materials such as samarium-cobalt (Sm-Co), neodymium-iron-boron (Nd-Fe-B). Alnico and ferrites have long been used in the development of PM motors as they are cheap and easily available. Rare-earth PM materials, namely SmCo, are used nowadays because of the high energy density caused by its high residual flux density, coercive force and low temperature coefficient. NdFeB is considered one of the best PM materials presently since it offers much higher residual flux density and coercive force. However, its only drawback is the temperature limit. Continuous efforts are being made to overcome this and it is hoped that this will enable PMBLDC motors to attain higher efficiencies and lower sizes along with other advantages.

PMBLDC motors may be classified into different categories such as number of phases, radial or axial field, cageless or with cage bars, surface mounted PMs or buried magnets, sinusoidal or rectangular fed motors etc. Some of them are briefly discussed in this section.

2.1 *Number of phases*

PMBLDC motors are developed in single phase in low power (< 50W) for tube axial fans to cool electronics equipments [25]. They are manufactured in two phase construction for home appliances such as solar PV fed refrigeration system, servo control etc. Most of the medium and high power rating motors are designed in three-phase construction similar to conventional *ac* motors. In some electric vehicles [22-24] and megawatt rating

iotors for submarine propulsion etc., designers have compelling reasons to increase the umber of phases to five, six or more in order to reduce the per phase power handling equirements.

.2 *Radial and axial field motors*

lost of the motors in the market are radial field type (cylindrical or salient pole construc- on). However, the axial field motors have some advantages over the conventional radial eld construction in terms of power density, torque to inertia ratio, peak torque, less magnet eight, low inductance, short winding turns, compact design etc. [6,8]. Axial field motors re designed in package, disk and sandwich type construction and have no iron in the rotor, esulting in low inertia. Axially directed magnetic field from rotor magnet interacts with idially directed currents in these axial field motors. The magnets are encapsulated in resin r plastic. Because of their construction, they are considered most suitable for robotics, omputer equipments, machine tools etc. [8].

Radial field motors are also designed with varying desired flux linkage waveforms such s sinusoidal or trapezoidal, different shapes and positions of magnets in the rotor such s buried or surface mounted etc. They are widely used since stator design is similar to onventional *ac* synchronous or induction motors. Figure 1 shows the typical cross-sections f these two types of popular PMBLDC motors.

.3 *Shape and location of PM in rotors*

'ermanent magnets are placed in the rotor in PMBLDC motors [6, 8]. In axial field type of notors, the magnets are encapsulated in resin or plastic in disc form as shown in figure 1a. 'hese magnets are placed in such a manner that induced back emf are either sinusoidal r trapezoidal waveforms. In radial field motors, the magnets are placed in different form uch as surface mounted for low speed motors and interior radially oriented or interior angentially oriented in high speed PMBL motors. Figure 2 shows such rotor geometries. 'hey are also designed to achieve sinusoidal or trapezoidal back emfs depending upon pplications.

.4 *Sinusoidal and rectangular fed motors*

'MBLDC motors are designed to have either sinusoidal or trapezoidal (excited) induced ack emfs [8,26–32]. Sinusoidal excited motors are fed with sinusoidal polyphase currents imilar to conventional synchronous motors for ripple-free torque with unity power factor or constant torque operation below base speed with frequency control and having lead- ng currents to affect field weakening for constant power operation. Maximum speed of peration is restricted with demagnetization caused by armature reaction and mechanical onstruction. Magnetic saliency on rotor with reluctance torque helps to achieve wide peed range of constant power operation. Trapezoidal excited motors need polyphase bal-

b

a
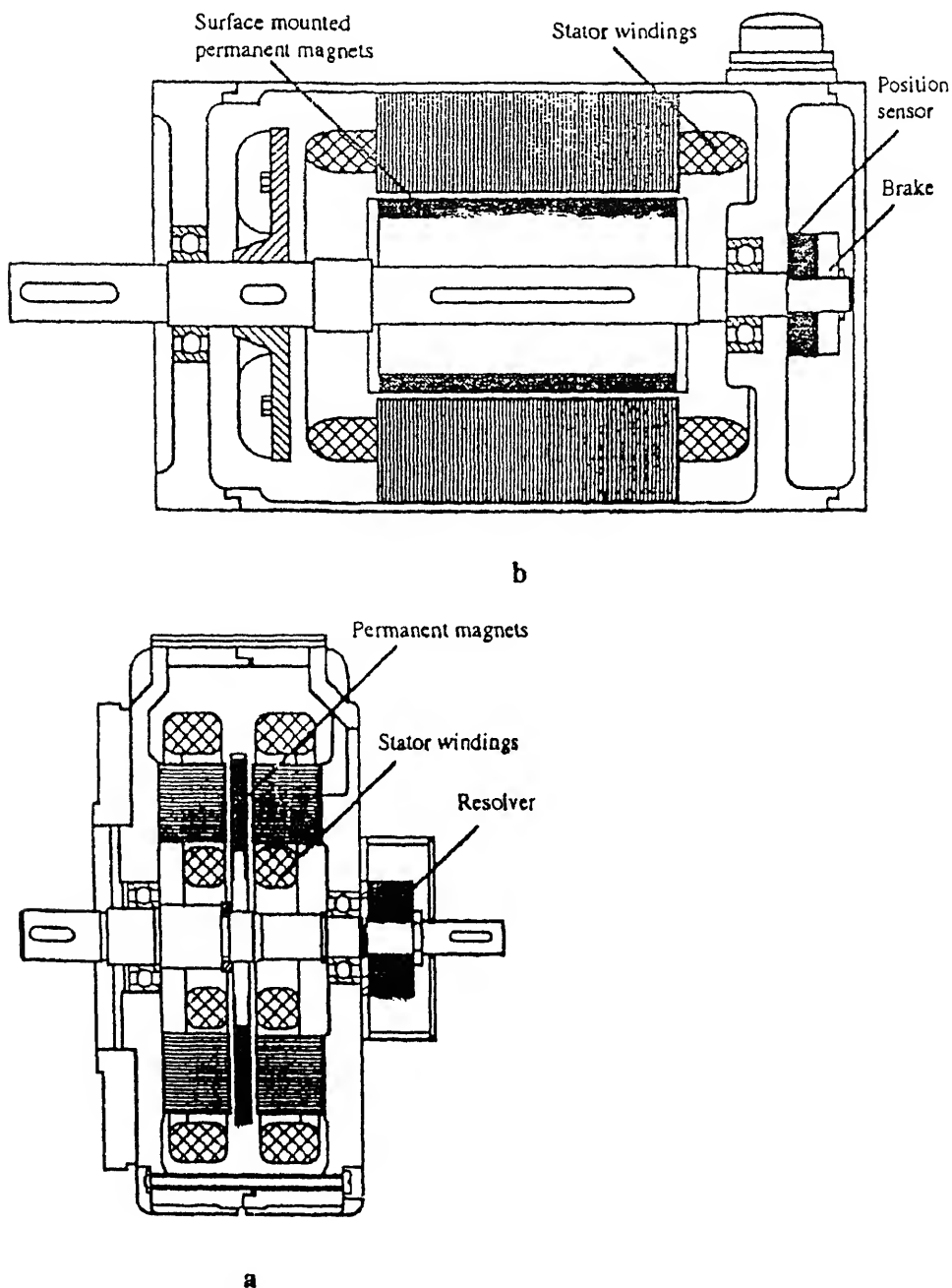
Figure 1.   Cross-sections of two types of PMBLDC motors (a) Axial field motor, (b) radial field motor.
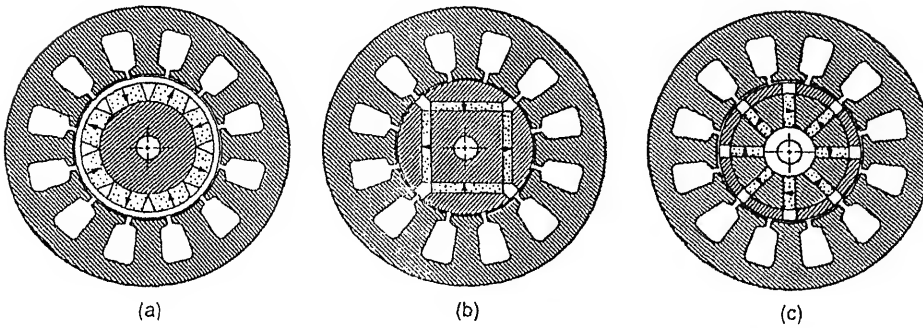
Figure 2. Rotor geometries of PMBLDC motors. (**a**) Surface mounted, (**b**) interior radially-oriented, and (**c**) interior tangentially oriented magnets.

these rectangular currents they are also called switched PM motors, brushless *dc* motors d electronic commutated PMBLDC motors.

Figure 3 shows the ideal current waveforms for these two types of motors. Position nsors requirement is accordingly changed to realize these ideal current waveforms in the otor windings in self synchronous control mode.

Closed loop controllers

respective of sinusoidal or trapezoidal excitation, PMBLDC motors are used for posi-n control, speed control and torque control in motion control applications [3,8]. Figure 4 ows a typical position closed loop control with inner speed and current loops. For speed ntrol system outer position loop is not required and speed reference is the command gnal. Torque control is incorporated in high performance motion control through closed op regulation of phase currents in synchronization with shaft position feedback. In the ajority of PMBLDC motors, torque is linearly related to currents and torque command aps into current commands with only a simple proportionality constant. In some typal cases, a nonlinear mapping is required between torque and current commands. The

(a) (b)
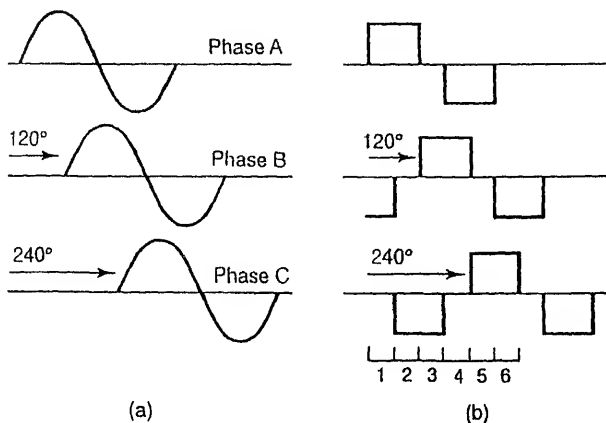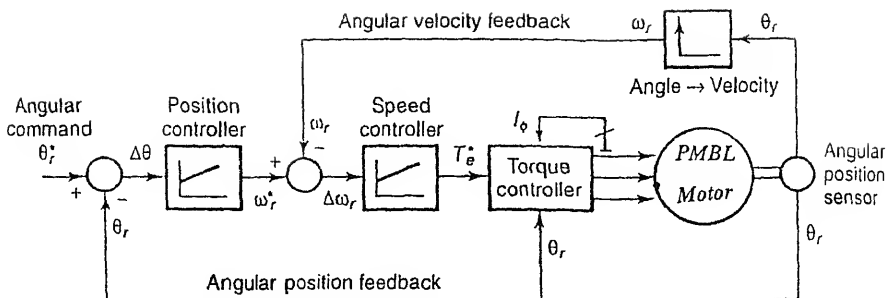
Figure 4. Block diagram of closed loop position control with speed and current loops of of PMBLDC motor drive.

constant power operation of PMBLDC motor drive is extended through field weakening control techniques which also requires another command signal for torque to current mapping. Current regulation in the phase windings of PMBLDC motors either in sinusoidal or rectangular manner is carried out using current controlled voltage source inverter(VSI). PWM, hysteresis and predictive current controllers are used to issue the switching signals to the devices of the inverter to realize winding currents close to command currents. Speed control is generally achieved by using a speed feedback and speed command through speed controller which outputs a command signal for torque controller. Position control is implemented through position feedback and position command using position controller. The output of the position controller is the speed command for inner speed loop. Both position and speed closed loop controllers are realized using wide varying closed loop controllers such as PI (proportional-integral), PID (proportional-integral-derivative) [8], SMC (sliding mode controller) [33–35], adaptive controllers [36], fuzzy based control [8,19] and neural based controllers [8].

These classical (PI or PID) controllers or advanced closed loop controllers such as SMC, fuzzy and neural network-based ones are implemented using DSP, microcontroller and specific application integrated chip (ASIC) for speed and/or position control. Many manufacturers have developed ASICs for typical applications of PMBLDC motors.

4. Recent developments in inverters and converters

PMBLDC motors are invariably fed from variable frequency inverters to provide electronic commutation. At small ratings, MOSFET-based VSIs are used to achieve ideal current control with reasonable high switching frequency. In medium rating drives, IGBT-based VSIs are used to feed PMBLDC motors. GTO-based inverters are used in high power rating PM-BLDCM drives because of their self-commutating feature and improved current control. In medium and high power rating PMBLDC motors, an input unity power factor rectifier with PWM current control is used to achieve regulated dc link and regenerative feature of the PMBLDC motor drive. Classical PI/PID and feedforward closed loop controllers are used for the control of front end converters. It improves the power quality of the ac mains in terms of reduced harmonics and unity power-factor in both the directions of power flow.

Recently MOSFET/IGBT-based VSIs are used in module forms to achieve compact sign of the drive. Power MOS ICs are used to control small rating PMBLDC motors ich has integrated microelectronics for control and power amplifier to excite the motor. reover, their gate drives are also used in module form with all protective and intelligent ntrol features. In some typical applications (residential commercial blower), they are de so compact that these are fitted inside the motor housing [8].

Latest trends in sensors

the control of PMBLDC motors, position, speed and current sensors are essentially re-ired to regulate the phase currents in synchronization with rotor position [3,4,8]. Moreover, metimes, terminal voltage sensors are also required to estimate either position or speed. ltage sensors are also needed to regulate *dc* bus voltage during braking or for front end nverter control. In some typical attempts, flux sensors and torque sensors are also used the precise control of these motors. Basic role of these sensors are already discussed in)sed loop control of PMBLDCM drive in § 3. In the following section, the recent trends sensors and their function are briefly discussed.

Position sensors

e rotor flux position in PMBLDC motors is defined by the mechanical angle of rotation, iich is achieved from some form of rotor position sensors. Rotor flux position is required phase current synchronization and rotor position is also required for position control. tor position is directly sensed using position sensors or indirectly estimated using other easured parameters. Hence, rotor position sensing is indispensable in current-controlled /IBLDC motor drives.

Rotor position is sensed using resolvers, inductive modular absolute system (IMAS) [37], ll effect position sensor, magnetoresistors [38], electronics and optical encoders, synchros d tachsyn [39]. The tachsyn is an airgap reluctance sensitive 3-phase alternator with PM ld and trapezoidal output waveforms. It is used for position and velocity sensing and gnal outputs are analog. They are available in 4, 6 and 8 pole configurations. The encoders e characterised by number of pulses per revolutions (PPR) and nowadays they are avail-le in several thousands PPR. Interfacing ICs are also available to convert these sensor gnals to digital form to feed digital processors used for intelligent control of PMBLDCM ives.

Indirect position sensing is achieved by estimating the rotor position using other mea-red parameters such as currents and voltages etc. There are many techniques for rotor ix position estimation which are much detailed in the next section.

Speed/velocity sensors

the PMBLDC motor drive, speed or velocity signals are essentially required for speed ntrol loops in position controlled drives and speed feedback for speed controlled drives.)eed measurement is carried out either using speed transducers/sensors or estimated

using the rotor position information either obtained through direct position sensing or through estimation.

In general *dc* tachogenerators and brushless tachogenerators are used to sense motor speed. They provide an analog *dc* voltage signal which is proportional to shaft speed. The polarity of this voltage signal in both types of tachogenerators results in the direction of rotation. Nowadays rotor velocity/speed is estimated more accurately by using high resolution position sensors or estimated rotor flux position. Sometimes, these sensors are different from the position sensors used for electronic commutation.

5.3 *Current sensors*

Fast torque control in high performance PMBLDC motor drives is implemented through closed loop regulation of phase winding currents in synchronization with rotor flux position information. Closed loop regulation of winding currents is realized through PWM or hysteresis current controllers of CC-VSI over the reference desired currents and sensed winding currents. Therefore, the sensing of winding currents becomes indispensable in PMBLDC motor drives.

The current sensing is generally carried out using hall effect current sensors. They detect the magnitude and direction of currents and are integrated to provide sensitive and accurate current sensing. Very fast response (less than 1 microsecond) and accurate current sensors are available from different manufacturers (ABB, LEM etc.) in wide range of current sensing (fractions of amperes to kiloamperes). These hall effect current transducers have galvanic isolation of several kilovolts which is a very desired requirement of these drives in high rating. Generally in 3-phase motors two current sensors are required and third-phase current is estimated from other two-phase currents in star connected motors. These current sensing requirements for current regulated rectangular fed PMBLDC motor drive are typically reduced to a single current sensor in the *dc* link of the inverter. Current shunt resistors with low power dissipation are often used as the current sensor in low power drives for cost effectiveness. In the modern power devices such as MOSFET/IGBT a current sensing feature is provided by many manufacturers which also dispenses with the use of extra current sensors in the control of the inverter/converter feeding the PMBLDCM.

5.4 *Voltage sensors*

Terminal voltage sensing in modern advanced PMBLDC motor drives is required to estimate the rotor position and speed for the control, resulting in mechanical sensorless drive with a view to reduce size, cost, maintenance and enhanced reliability. Voltage sensing is also required to regulate *dc* bus voltage during braking or for control of front end converter used for regenerative feature in high rating drives.

Terminal voltage sensing is carried out by using electronic isolation amplifier (AD202 Analog Devices make etc.) and hall effect voltage sensors (ABB make etc.) with galvanic isolation. In small rating drives, voltage is sensed using high valued resistor potential dividers to reduce the cost of the drive. Sometimes, induced voltage in the motor windings is achieved using special windings such as search coils etc. However, *ac* mains voltages for the control of front end converter are sensed using the potential transformers.

Sensors elimination and reduction

cent trends in sensors, their requirement and types of available sensors are already
cussed in § 5. However, some of these sensors in PMBLDC motor drive may be reduced
m the view point of size, cost, maintenance and reliability. Typically mechanical rotor
sition and speed sensors have the drawbacks of increasing the number of connections
ween motor and controller, increased interference, limitation in accuracy of sensors
e to environmental factors such as temperature, humidity, vibrations etc., increased
ction and inertia and additional space in motor housing. Because of these problems
ently there has been wide interest and developments in the techniques for elimination
mechanical rotor position/speed sensing by estimating the rotor position and speed
ng sensed currents and voltages [40–57]. Moreover, the number of voltage and current
sors may be reduced through using intelligent processors for the control of inverter
ding PMBLDC motors. Various techniques for these sensors elimination and reduction
briefly discussed in the following section.

Mechanical sensors elimination

e of the most recent developments in PMBLDC motor drives has been the rapid evolution
new techniques for eliminating the rotor angular position sensor. Elimination of the
ft-mounted position sensor is a very desirable feature in a number of applications since
s sensor is one of the most expensive and fragile components in this drive. Some of the
sition sensorless schemes are classified in brief.

Back EMF based position estimation

e most common methods of rotor flux position sensing is based on deriving the back
if signals. There are many methods for rotor position estimation based on the back emf
d are briefly discussed below.

a Direct back EMF detection: This method is quite popular for rectangular fed
MBLDC motors. In these PM motors, particular phase winding is excited for 2/3 of each
ectrical period and ideally there is always one phase which is not excited. The direct
nsing of back emf of unexcited phases in sequence is used to generate discrete rotor
sition signal for current synchronization with rotor flux. It has been applied to many
dustrial applications including disk drives [46], compact stereo players [41,42] and room air
nditioners [43].

b Estimation of back EMF: This method is applied to both sinusoidal fed and rect-
gular fed PMBLDC motors. This method is based on the reconstruction of the back
if by using voltage equation of the motor ($e = v - iR - L \mathrm{d}i/\mathrm{d}t$). Reconstruction of the

terminal voltage and line current are measured directly and the above equation is used to achieve back emf and rotor position.

6.3   *Third harmonic voltage detection based position estimation*

In star-connected PMBLDC motors, the third harmonic voltage is measured between the star point and an artificial star point created by three high-value resistors which are connected to the motor terminals. Such a voltage gives six zero crossings in the 3-phase motor and results in rotor position for current synchronization with rotor flux.

6.4   *By monitoring current or computation of the phase inductance*

The basic concept of this method is that the rate change of current in a phase winding of the motor depends on the incremental inductance which is rotor position dependent. This phase winding inductance variation with rotor position is used to estimate the rotor position for electronic commutation of the inverter. This current sensing model for rotor position estimation is used by Lin *et al* [50].

6.5   *Injecting diagnosis signal to the stator winding*

This method uses a PWM carrier frequency and inductance bridge to measure the rotor flux path reluctance. The method operates on a bridge principle by monitoring the inductance difference in two phases, and is sensitive to small variations in reluctance.

6.6   *Observer methods*

In these methods, an observer reconstructs the rotor position which is directly measurable. Basically all these methods use the sensed phase and/or line currents to perform on-line compensation to derive the rotor position. A number of observer methods such as Kalman filter technique, discrete time observer, state observer and stator flux estimation method current and the voltage based observer method [56–57] as well as improved different types of motors, are reported.

6.7   *By a special windings electromagnetic devices*

A number of methods based on special winding such as search coils or an electromagnetic device [58] are used to sense the rotor position. The electromagnetic device consists of pick-up coils around a special stator made of magnetically nonlinear material. The pick-up coils are excited by a high frequency sinusoidal current. The device detects the phase of the second harmonics component of the induced voltage in the pick-up coils.

6.8   *Monitoring switching states in the inverter*

This method is applied to rectangular fed PMBLC motor and ON/OFF states of inverter switching devices are used for rotor position estimation [51]. The method is based on a

or with trapezoidal back-emf. The rotor flux position is obtained on the basis of the ducting state of free-wheeling diodes in an open phase at a particular time.

n most of these methods, the rotor speed/velocity is estimated by using the time deriva- of rotor position angle or by measurement of the period during the transition of alter- ng rotor position.

Elimination/reduction of current and voltage sensors

nsive work is carried out to reduce the current and voltage sensors in the PMBLC or drive to reduce the cost and enhance reliability. Normally two current sensors are uired in 3-phase star connected motors, a technique used to estimate the 3-phase winding rents only using one current sensor in the *dc* link and switching states of the inverter ices. Since, in the intelligent inverter control, device-switching patterns are available he processor, the 3-phase winding currents signals are constructed using measured *dc* current and switching status of the inverter devices. Similarly, one voltage sensor is d at the *dc* link and three-phase terminal voltages are derived using the same switching es of the inverter devices.

'or rectangular fed PMBLDC motor drives, current sensors can be eliminated entirely g current sensors embedded in three of the six inverter switches [59]. MOS gated devices h as MOSFET and IGBTs in integrated modules from different manufacturers are ilable which incorporate current sensors integrated into monolithic power devices.

Torque pulsations, noise, vibration and their reduction

y divergence from ideal conditions either in the motor (design factors) or in the power erter feeding PMBLDC motor drive (current waveforms) results in undesired torque sations [60–68]. The torque pulsation in PMBLDCM drive causes speed oscillations, itation of resonances in mechanical portions of the drive causing acoustic noise and ible vibration patterns in high precision machines. Pulsating torque in PMBLDC motors asically in form cogging torque generated by interaction of the rotor magnetic flux with or magnetic reluctance variation and ripple torque generated by the interaction of stator rent mmfs with rotor magnets. Ripple torque is due to mutual or alignment torque caused current mmf with rotor flux distribution and reluctance torque caused by current mmf h rotor reluctance variations. In rectangular fed PMBLDC motor, the fringing fields at rotor pole edges cause deviations from ideal trapezoidal emf waveform and currents also not strictly rectangular resulting in dip of the torque magnitude up to 25% of rated que [67].

Torque pulsations are minimized by two broadly classified techniques, namely improved tor designs and active control schemes. Improved design techniques for pulsating torque nimization include skewing, fractional slot winding, short pitch winding, increased nber of phases, airgap windings, adjusting stator slot opening and wedges [68], rotor

However, active control techniques for pulsating torque minimization include adaptive control technique [8,67], preprogrammed current waveform control, selective harmonics injection techniques, estimators and observers, speed loop disturbance rejection, high speed current regulators, commutation torque minimizations [67,68] and automated self commissioning schemes [67].

8.   Field weakening

Field weakening of PMBLDC motors has become quite demanding for achieving high speed operation in constant power applications such as electric vehicles etc. [8,69]. For speeds above base speed, the rotational emf increases rapidly and constant power operation cannot be maintained at high speeds. The amplitude of back emf (line to line) increases linearly with speed and becomes more than the *dc* link voltage of the inverter. Eventually the current controller saturates losing its ability to force the desired reference currents into the motor phases.

Reasonable torque production may be attained by advancing the phase angle of the excitation transition points relative to back emf. As this angle is advanced, current in on coming phase is given a controlled time interval to build up before the back emf increases and chokes off further current growth. This concept is illustrated through figure 5. Figure 5a shows the concept of advance angle (α) in back emf and current of a particular phase winding of rectangular fed PM motor for 2/3 period excitations (120°). However, this concept is also applicable for current excitation of 180°. The resulting effects of this advance angle excitation on torque-speed characteristics of the PMBLDCM are shown in figure 5b. This concept is used in both sinusoidal and rectangular fed motors to extend the speed range as constant power drive. This control aspect can give at the most, the speed range 3 : 1. This speed range may only be extended to 7 : 1 or more by proper designing of the motor such as axially laminated PM reluctance motor.
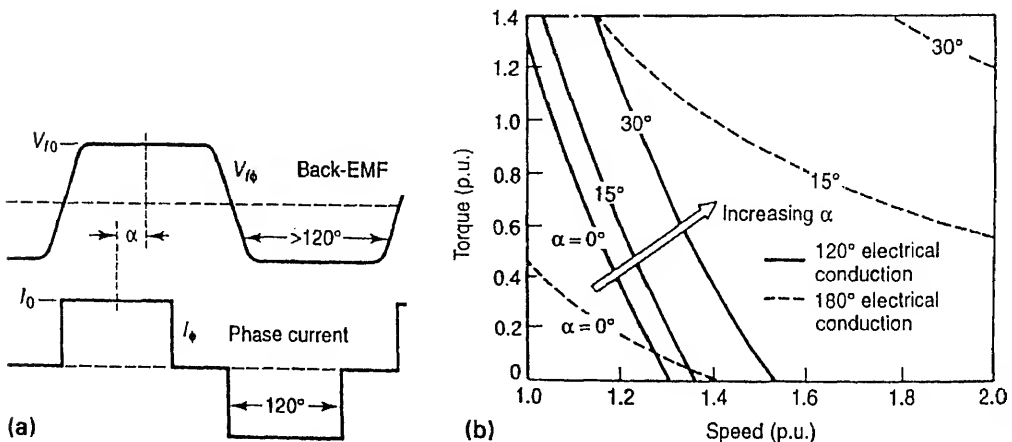
Figure 5.   (a) Typical back *emf* and phase current waveforms of rectangular fed PMBLDC motor drive with advance angle. (b) Typical torque-speed curves of rectangular fed PMBLDC motor drive with advance angle control.

9. Impact of microelectronics on PMBLDC motor drive

Improved and intelligent control techniques for the PMBLDCM drive are implemented through advanced technology in integrated electronic circuit chips [70]. ASICs (specific application integrated chips) are developed by many manufacturers namely LM621 by National Semiconductors Limited, (Si9985CY) by Silconix, SA/SE/NE5570 by Phillips etc. for the PMBLDCM drives. Moreover, sensorless operation of PMBLDC motors, adaptive control, torque pulsation minimization etc are realized through micro controllers, DSP and single chip microcomputers. Reduced cost of manufacturing of VLSI has made it possible to design dedicated ASICs for such complex and intelligent control and providing compact, reliable, low cost and effective controller for PMBLDCM drives [70]. For many dedicated applications, power and control are integrated in the same MOS ICs for low power PMBLDCM drives.

10. Application issues of PMBLDC drives

PMBLDC motors are attractive in drive applications which benefit from high efficiencies as compared to induction motor and switched reluctance motors. Sinusoidal-fed PMBLDC motors are considered superior to rectangular-fed PMBLDC motors for achieving the wide high speed constant power speed range [8]. The positive characteristics of PMBLDC motor discussed earlier make them highly attractive for a number of applications, a few of which are given here.

10.1 *Servo actuators*

High power density, fast dynamic response, smallest size and weight with rare earth magnet (Nd Fe-B or SmCo) of PMBLDC motors have made them the best suited candidates in machine tool servos, robotics actuators drives and dual tandem electromagnetic actuators (EHA) [21].

10.2 *Commercial residential speed control applications*

PMBLDC motors drives are used in a wide range of commercial and residential applications such as domestic appliances, heating, ventilating and air conditioning equipment due to their highest possible efficiencies. The speed control ability of compressors and blowers is able to provide operation at their high efficiency. The physical integration of controller electronics in the motor body itself is able to make them most suitable for low power (0.5 hp) blowers and low power (50 W) tube axial fans for cooling the electronics equipment [25].

10.3 *Automotive applications*

The efficient variable speed feature with low volume of PMBLDC motors has made them

capable of providing wide speed range constant power characteristics at high efficiency to match these automotive applications.

10.4 *High power industrial and propulsion drives*

High power (1.1 MW), six-phase PMBLDC motors are considered best-suited for direct drive marine propulsion application to replace the *dc* motor drive with a saving of 40% in weight, length and volume. PMBLDC motors upto 2 MW power rating and speed of 4000 rpm are used in land and marine propulsion systems and industrial pumping.

With high efficiency and high power density, the demand for PMBLDC motor drives is increasing and is expected to continue for high performance servos, commercial residential products, computer industry, information technology, propulsion systems, vehicle and traction.

11. Conclusions

The latest trend and present status of PMBLDC motor drives have revealed that this drive has great potential for use in a number of applications. The use of new PM materials, latest motor construction technology, intelligent closed-loop control, new types of inverters, sensorless operation, and field weakening have resulted in potential applications and provided a broad prospective to drive-design and application engineers. The methods of torque pulsation reduction have been explored to design PMBLDCM drives to suit highly position-sensitive applications. In view of these new developments, these drives are expected to enjoy a bright future in a larger number of applications of adjustable speed applications requiring high efficiencies and fast dynamic responses such as blowers, compressors, automobiles, vehicle propulsion systems and traction drives.

References

[1] T Kenjo, S Nagamori 1985 *Permanent magnet brushless DC motors* (Oxford: Clarendon)

[2] T J E Miller 1989 Brushless permanent magnet and reluctance motor drive (Oxford: Clarendon)

[3] P Pillay (ed.) 1989 Performance and design of permanent magnet AC motor drives. *IEEE Industry Applications Society Tutorial Course*, IEEE Industry Applications Society Meeting, San-Diego, CA

[4] Y Dote, S Kinoshita 1990 Brushless servomotors fundamentäls and applications. Oxford: Clarendon

[5] Y Dote 1990 *Servo motor and motion control using digital signal processors.* (Englewood, Cliffs, NJ: Prentice Hall)

[6] D C Hanselman 1994 *Brushless permanent magnet motor design* (New York: McGraw-Hill)

[7] J F Gieras 1996 Permanent magnet motor drives. *Power Electronics, Drives & Energy Systems'96*, Tutorial, Indian Inst. Technol., Delhi

[8] B K Bose (ed) 1997 *Power electronics and variable frequency drives – Technology and applications.* (New York: IEEE Press)

[9] D Weinmann, G Nicoud, F Gallo 1984 Advantages of permanent magnet motors. *Proc. of Drives/Motors/Controls-84*, Brighton, UK, pp 113–120

[10] L M C Mhango 1989 Benefits of Nd-Fe-B magnet in brushless DC motor design for aircraft applications. *Proc. of 4th International Conference on Electrical Machines and Drives*, IEE Conf. Pub. No. 310, pp 76–79

[11] K J Binns 1994 Permanent magnet drives; the state of the art. *Symposium on Power Electronics, Electrical Drives, Advanced Electric Motor SPEEDAM 94*, Taormina, Italy, pp 13–18

[12] M A Rahman, G R Slemon 1985 Promising applications of neodymium, boron and iron magnets in electrical machines. *IEEE Trans. Magne.* MAG-21: 1712–1716

[13] E Ritchter, T J E Miller, T W Neumann, T L Hudson 1985 The ferrite permanent magnet AC motors – A technical and economical assessment. *IEEE Trans. Ind. Appl.* IA-21: 644–650

[14] R Hanitsch, C S Park 1990 Novel 10 W brushless DC motor of the Pankake type (IEE CD No.-324) *4th International Conference on Power Electronics and Variable Speed Drives*, pp 435–439

[15] D Pauly, G Plaff, A Weschta 1984 Brushless servo drives with permanent magnet motors or squirrel cage induction motors – A comparison. *IEEE IAS Annual Meeting*, pp 503–508

[16] D Howe, M K Jenkins, Z Q Zhu 1993 Permanent magnet machines and drives – An integrated design approach. (IEE CP No. 376) *IEE Sixth International Conference on Electrical Machines and Drives* Oxford, UK, pp 625–630

[17] T S Low, K J Binns 1986 Multistacked imbricated rotors with permanent magnet excitation; Design for new magnetic materials. *Inst. Elec. Eng. Proc.* B133: 205–211

[18] S Williams 1985 Direct drive system for an industrial robot using a brushless DC motor. *Inst. Elec. Eng. Proc.* B132: 53–56

[19] M A Jabbar 1996 Disk drive spindle motors and their controls. *IEEE Trans. Ind. Electron.* 43: 276–284

[20] W L Soong, D A Staton, T J E Miller 1993 Design of a new axially-laminated interior PM motor. *IEEE-IAS Annual Meeting Record*, pp 27–36

[21] T M Jahns, R C Van Nocker 1990 High performance EHA control using an interior PM motor. *IEEE Trans. Aero. Elect. Syst.* 26: 534–542

[22] B Sneyers, G Maggetto, J Van Eck 1992 Inveter fed permanent magnet motor for road electric traction. *Proc. of ICEM-1992*, Budapest, pp 550–553

[23] A R Millner 1994 Multi-hundred horsepower permanent magnet brushless disk motors. *Proc. of APEC Conf.*, pp 351–355

[24] C C Chen, K T Chau, J Z Jiang, W Xia, M Zhu, R Zhang 1996 Novel permanent magnet motor drives for electric vehicles. *IEEE Trans. Ind. Electron.* 43: 331–339

[25] L Newborough 1990 Electronically commutated DC motor for driving tube axial fans: A cost effective design. *Appl. Energy* 36: 167–190

[26] G Liu, W G Dunford 1990 Comparison of sinusoidal excitation and trapezoidal excitation of a brushless permanent magnet motor. (IEE CD No 324) *4th International Conference on Power Electronics and Variable Speed Drives*, London, pp 446–450

[27] M Allan, I J Kemp, 1993 Commutation strategies for the DC brushless motor. (CP No. 376) *Sixth International Conference on Electrical Machines on Drives*, Oxford, UK, 133–178

[28] J De La Ree 1990 Performance evaluation of PM machines with quasi-square wave input currents. *Elec. Mach. Power Syst.* 18: 283–291

[29] C S Berendsen, G Ckampenois, A Bolopion 1993 Commutation strategies for brushless DC motors: Influence on instant torque. *IEEE Trans. Power Electron.* 8: 231–236

[30] A Rubaai, R C Yalmanchili 1992 Dynamic study of an electronically brushless DC machine via computer simulations. *IEEE Trans. Energy Conversion* 7: 132–136

[31] M T Wishart, R G Harley, C Diana 1991 The application of field oriented control to the brushless DC machine. *Proc. Euro. Power Electron. Conf.*, Ferenze, pp 629–634

[32] C L Putta Swamy, B Singh, Bhim Singh 1995 Investigations on dynamic behavior of permanent magnet brushless DC motor drive. *J. Elec. Mach. Power Syst.* 23: 689–701

[33] C Rossi, A Tonielli 1994 Robust control of permanent magnet motors: VSS techniques lead to simple hardware implementations. *IEEE Trans. Ind. Electron.* 41: 451–460

[34] G Carrara, D Casini, A Landi, L Taponecco 1991 Sliding mode speed controller for trapezoidal brushless motors. *Elec. Mach. Power Syst.* 19: 157–169

[35] B Singh, B P Singh, C L Putta Swamy 1995 Modeling of variable structure controlled permanent magnet brushless DC motor: *J. Inst. Eng. (India)* 75: 183–189

[36] E Cerruto, A Consoli, A Raciti, A Testa 1995 A robust adoptive controller for PM motor drives in robotic applications. *IEEE Trans. Power Electron.* 10: 62–71

[37] G R Horner, W Freund 1991 A new approach to multi-turn absolute position, velocity and motor commutation signals. *Proc. of Drives/Motors/Controls 1991*, pp 155–158

[38] C Ferreira, D Belanger, J Vaidya 1987 A magnetic rotor position sensor for brushless permanent magnet motors. *Proc. of MOTOR-CON 1987*, pp 146–156

[39] J R Luneau 1985 New developments in feedback devices for brushless DC servosystems. *Proc. of MOTOR-CON 1985*, pp 86–95

[40] R Krishnan, R Ghosh 1987 Starting algorithm and performance of a permanent magnet brushless motor drive with no position sensor. *IEEE Power Electron. Syst. Conf. 1987*, pp 596–606

[41] N Ertugrul, P Acarnley 1994 A new algorithm for sensorless operation of permanent magnet motors. *IEEE Trans. Ind. Appl.* 30: 126–133

[42] P P Acarnley, N Ertugrul 1992 Rotor position estimation in PM motors. *International Conference on Electrical Machines*, pp 1–5

[43] T Endo, F Tajima, H Okuda, 1983 Microcomputer-controlled brushless motor without a shaft-mounted position sensor. *International Power Electronics Conference, IPEC*, Tokyo, pp 1478–1488

[44] H Hzuka, H Uzuhashi, M Kano, I Endo, K Mohri 1985 Microcomputer control for sensorless brushless motor. *IEEE Trans. Ind. Appl.* IA-21: 595–601

[45] N Matsui, M Shigyo 1992 Brushless DC motor control without position and speed sensors. *IEEE Trans. Ind. Appl.* 20: 339–346

[46] B C Kuo, K Butts 1982 Closed loop control of a 3.6 degree floppy-disk drive PM motor by back EMF sensing. *11th Proc. of SIMCSO*, Champaign

[47] J Hu, D M Dawson, K Anderson 1995 Position control of a brushless DC motor without velocity measurements. *IEE Proc. Elec. Power Appl.* 142: 113–122

[48] K J Binns, D W Shimmin, K M Al-Aubidy 1991 Implicit rotor position sensing using motor windings for self-commutating permanent magnet drive system. *Inst. Elec. Eng. Proc.* B138: 28–34

[49] T Furuhashi, S Sangwongwanich, S Okuma 1992 A position and velocity sensorless control for brushless DC motors using an adaptive sliding mode observer. *IEEE Trans. Ind. Elec.* 39: 89-95

[50] R L Lin, M T Hu, C Y Lee 1989 Using phase current sensing circuit as the position sensor for brushless DC motor without shaft position sensor. *Proc. of IEEE-IECON-1989*, Part 1

[51] S Ogasawara, H Akagi 1991 An approach to position sensorless drive for brushless DC motors. IEEE Trans. Ind. Appl. 27: 000-000

] H Watanabe, T Ishii, Fujii DC brushless servo system without rotor position and speed sensor. *Proc. IEEE-IECON-1987*, Cambridge, MA

] R Wu, G R Slemon 1991 A permanent magnet motor drive without a shaft sensor. *IEEE Trans. Ind. Appl.* 27: 00-00

] P W Lee, C Pollock 1992 Rotor position detection techniques for brushless PM and reluctance motor drives. *IEEE-IAS Annual Meeting Record*, pp 448-455

] R C Becerra, T M Jahns, M Ehsani 1991 Four quadrant sensorless brushless ECM drive. . *Proc. of Appl. Power Electron. Conf.*, pp 202–209

] N Matsui, M Shigyo 1990 Brushless DC motor control without position and speed sensors. *IEEE-IAS Annual Meeting Record*, pp 448–453

] N Matsui 1996 Sensorless PM brushless DC motor drives. *IEEE Trans. Ind. Electron.* 43: 300–308

] D E Hesmondhalgh, D Tipping 1990 An electromagnetic motor integrated position sensor for brushless DC motors, capable of operation at standstill. *Proc. of ICEM-1990*

] T M Jahns, R C Becerra, M Ehsani 1990 Integrated current regulation for brushless ECM drives. *IEEE Trans. Power Electron.* 6: 118–126

] B Ackermann, J H H Janssen, R Sottek, R I Van Steen 1992 New technique for reducing cogging torque in a class of brushless DC motors. *Inst. Elec. Eng. Proc.* B139: 00-00

] H Bolton, R Ashen 1984 Influence of motor design and feed current waveform on torque ripple in brushless DC drives. *Inst. Elec. Eng. Proc.* B131: 82–90

] H Le-Huy, R Perret, R Feuillet 1986 Minimization of torque ripples in brushless DC motor drives. *IEEE Trans. Ind. Appl.* IA-22: 748–755

] Y Murai, Y Kawase, K Ohashi, K Nagatake, K Okuyama 1987 Torque ripple improvement for brushless DC miniature motors. *IEEE-IAS Annual Meeting Record*

] J Y Hung, Z Ding 1993 Design of currents to reduce torque ripple in brushless permanent magnet motors. *Inst. Elec. Eng. Proc.* B140: 260–266

] T Li, G Slemon 1988 Reduction of cogging torque in permanent magnet motors. *IEEE Trans. Magn.* 24: 2901–2903

] F Leonardi, M Venturuni, A Vishmara 1994 Design and optimization of very high torque, low ripple, low cogging PM motors for direct driving optical telescopes. *IEEE-IAS Annual Meeting Record*

] J Holtz, L Springgob 1996 Identification and compensation of torque ripple in high precision permanent magnet motor drives. *IEEE Trans. Ind. Electron.* 43: 309–320

] T M Jahns, W L Soong 1996 Pulsating torque minimization techniques for permanent magnet AC motor drives – A review. *IEEE Trans. Ind. Electron.* 43: 321–330

] B Sneyers, D W Novotny, T A Lipo 1985 Field weakening in buried magnet AC motor drives. *IEEE Trans. Ind. Appl.* IA-21: 398–407

] D Kinniment, P Acarnley, A Jack 1991 An integrated circuit controller for brushless DC drives. *Proc. of EPE*, Florence, 4: 111–116

C motor traction drives – A status review

L FREDERICK[1] and GOPAL K DUBEY[2]

[1]Transportation Division, Unit III, Kirloskar Electric Co. Ltd.,
Bangalore 560 058, India
[2]Department of Electrical Engineering, Indian Institute of Technology,
Kanpur 208 016, India
e-mail: gdubey@iitk.ernet.in

Abstract. The paper presents a brief review of modern *ac* motor traction drives, powered from 25 kV power frequency *ac* supply. Drives employing voltage source inverters, current source inverters, and induction and synchronous motors have been discussed.

Keywords. Traction converters; electric traction drives.

Introduction

wing to the advantages associated with *ac* motors, *ac* motor drives are fast replacg *dc* motor drives. Modern *ac* motor traction drives have several important features ch as regenerative braking capability, very little maintenance and down-time, nearnity *pf*, nearly sinusoidal current, good adhesion, high average speed and high fficiency.

Important features of electric traction drive (Dubey 1994, p. 327)

.) For economic reasons, a 25 kV, 50 Hz, *ac* traction drive employs only single-phase supply although the rating of the locomotive is as high as 6000 hp.

2) The traction supply system is weak in nature and hence the voltage is subject to fluctuations.

3) As traction supply is weak in nature, the reactive power has a very adverse effect. It is essential that the power factor is not allowed to be lower than 0.8.

4) Use of electric brakes reduces wear and tear on the track, wheels and brake shoes, thereby increasing their life substantially.

(6) Regenerative braking is generally combined with the dynamic braking.

(7) Use of the controlled rectifiers results in the generation of harmonics, which are injected into the source. This results in the following adverse effects.

    (a) High frequency harmonics cause interference with the communication line.

    (b) Low frequency harmonics may enter into the track circuits, leading to the mal-operation of the signals.

    (c) They can cause sharp fluctuations in the supply voltage.

3. Advantages of AC motor traction drives over DC motor traction drives

Modern traction drives generally use three-phase induction motors as traction motors. The *ac* motor drive has the following advantages.

    (i) Three-phase induction motors are robust and have high torque-to-weight ratio.

    (ii) They have simplified and reduced maintenance because of the absence of the commutator and brush gear as in *dc* motor drives.

    (iii) Full use of the available adhesion between wheel and rail is possible because of the naturally steep torque-speed characteristic of the induction motor.

    (iv) Induction motors have a good regeneration capability.

    (v) Due to absence of a commutator, the motor windings can be designed for higher voltages. This results in more favourable design of other components such as inverters, converters, transformer secondaries etc.

    (vi) Current rating of the induction motor is low. This results in reduction of cable size, number of contactors, number of switches etc.

    (vii) At the same power levels, induction motors are lighter than *dc* motors. This results in the relatively smaller unsprung mass of the truck giving good riding characteristics and low rail stress.

    (viii) Three phase induction motors have improved efficiencies and reliabilities in operation than do *dc* motors.

    (ix) With the three-phase drive, electric braking down to standstill is possible.

4. Suitability of VSI and CSI drive for traction application (Dubey 1997)

A three-phase AC induction motor should be fed from a three-phase supply capable of delivering steplessly variable frequency and voltage. The current source inverter and the voltage source inverter are available to meet this requirement. The important relative merits and demerits of CSI and VSI drives in relation to traction are as follows.

    (i) CSI drives do not suffer from the shoot-through fault which is common in VSI drives.

    (ii) Regenerative braking capability is inherent in CSI drives fed from a line-commutated fully controlled converter or a PWM fully controlled converter and in VSI drives

powered from a synchronous link converter. If the *ac* supply fails, the regenerative braking will not be possible in both the drives. Under such conditions, a VSI drive can use dynamic braking but the same is not possible with a CSI drive.

) The presence of a large value inductance in the *dc* link of a CSI drive results in the slower dynamic response compared to a PWM VSI drive. Consequently VSI drive has better adhesion (i.e. lower possibility of wheel slip).

) When the source is *dc*, a PWM VSI drive will be cheaper compared to a CSI drive of the same rating. Also the requirement of the large commutation capacitors, and a large *dc* link inductor (which is over-sized to prevent saturation), the volume and weight of a CSI drive is much higher compared to PWM VSI drive.

) The frequency range of the CSI is lower than that of the VSI drive. Hence the CSI drive has a lower speed range.

) The CSI is not suitable for multimotor drives. Hence each motor is fed by its own inverter and rectifier. But a single diode bridge or a synchronous link converter can be used to feed a number of VSI motor systems. Alternatively a single VSI can feed a number of motors.

AC motor traction drive

Principle of operation

e block diagram of a popular *ac* motor traction drive is shown in figure 1.
The pantograph collects the power for the running locomotive from the overhead line.
e pantograph is connected to the primary of the transformer. The isolation and protection vices are provided between the pantograph and transformer.
The traction transformer has a single primary winding and multiple secondary windings feeding traction converters. The traction transformer is so designed that the percentage
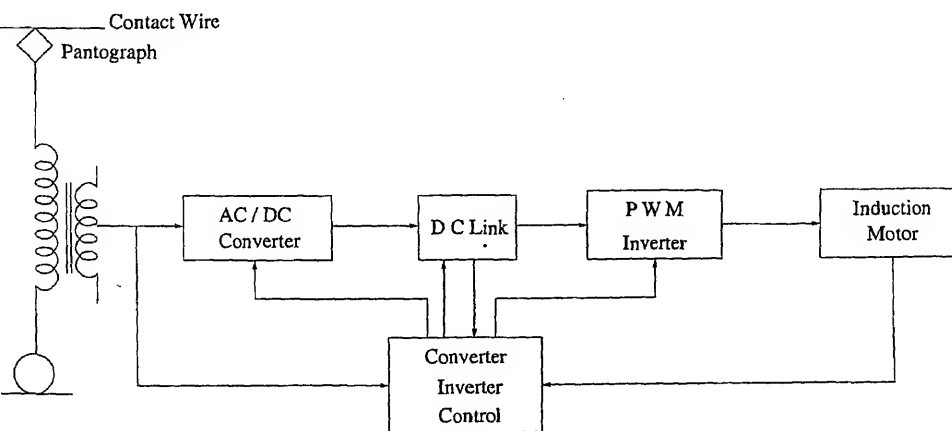
Figure 1. Block diagram of a popular *ac* motor traction drive.

impedance of all the secondary windings feeding traction converters are maintained un-changed irrespective of the combined or isolated operation of the windings.

Traction power converters consist of PWM converters. These converters are connected in series or parallel. Each unit is of modular construction with built-in cooling arrangement. Converters are operated in a way so as to maintain a power factor of unity throughout the operating range, with practically zero harmonic injection into the line.

The dc link consists of a suitable bank of capacitors designed to provide a stable dc link voltage for feeding variable voltage variable frequency (VVVF) inverter in voltage source configuration. Alternatively dc link may consist of a suitable inductor designed to provide a stable dc link current for feeding CSI inverter in current source configuration.

The dc link decouples the drive from the source. The inverter used may be CSI or VSI type. Generally CSI type inverter is used with synchronous motor drives and VSI type inverter is used with induction motor drives. The output of the inverter is connected to three-phase traction motors. The synchronous motors have higher full load efficiencies and power factors than induction motors. However, compared to squirrel cage induction motors they have higher cost, weight and volume for the same rating and require more maintenance.

Generally, microprocessor-based control systems are used for control of converters, dc links, inverters, traction motors, braking and slip. The microprocessor also performs the task of fault diagnosis and display in addition to that of control.

5.2 *Control of three-phase ac motor traction drive*

Several drives employing squirrel cage induction motors and synchronous motors are in use for traction. Variable frequency control is used both for induction motors and synchronous motors.

Modes of operation of an induction motor traction drive are shown in figure 2. This drive has received wide acceptance.
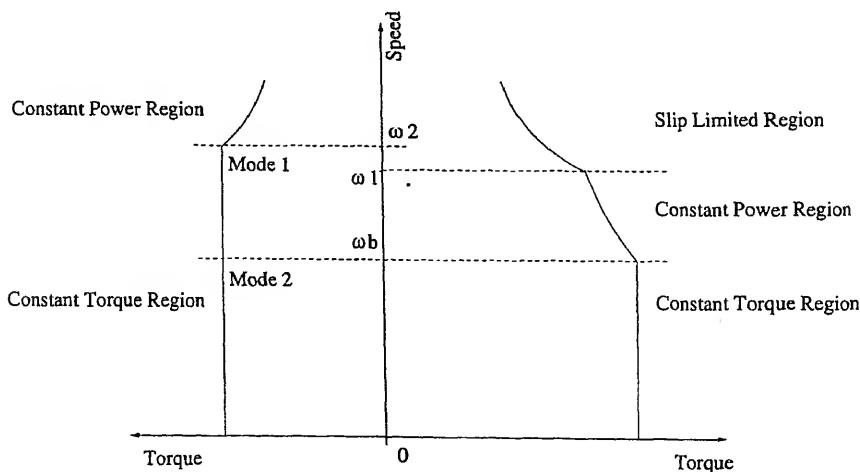
Figure 2. Modes of operation of ac motors with VVVF control.

/hen the drive is in the motoring region, the characteristic curve has three regions
lenkamp & Peak 1985; Dubey 1989).

Constant torque region;

constant power region;

slip limited region.

Constant torque region: The constant torque region ranges from standstill to base
ed (ω_b). In this region the inverter is operated to supply adjustable voltage and frequency
ne motor. The motor voltage is adjusted as speed (frequency) changes to maintain a
stant flux density in the motor. The motor voltage therefore increases proportionally
speed (frequency). The frequency of the voltage induced in the rotor (slip frequency)
eld constant. This produces nearly constant torque and, since the power output is
oortional to the speed, the power increases linearly with speed. At base speed (ω_b), the
/oltage reaches the maximum limit of the inverter.

Constant power region: In this region, the inverter is operated to supply the adjustable
quency to the motor. The voltage is no longer adjustable and the inverter is operated
naximum voltage limit. The power output of the motor remains nearly constant. This
le of control is extended until point ω_1, where the motor breakdown torque limit is
:hed.

Slip limited region: In this region, the slip frequency is maintained constant and
increase in speed is achieved by reducing the motor current. Here the motor current
uces inversely with speed and the torque decreases inversely as speed squared. This
racteristic is often referred to as the series motor characteristic. The torque produced
his region is somewhat higher than that produced by a *dc* series motor.
When the drive is in the braking region, the characteristic curve has the following regions
operation (Plunkett & Plette 1977).

Constant power region;

constant torque region.

Constant power region: In this region, drive power is held constant to match the
erter maximum power capability. This mode is similar to the constant power mode
ing motoring and slip frequency varies in proportion to the speed (frequency).

Constant torque region: When the drive is braking, the constant torque region has
modes of operation.

Model 1 – In this region, both the motor voltage and current vary approximately as the
are root of speed. Hence the power varies directly with speed, as the slip frequency is

Table 1.   Enforced limits, for psophometric currents.

Particular	Enforced limits
Psophometric current	$\leq$10 A
DC component of current	$\leq$4.7 A
Second harmonic current	$\leq$8.7 A
Audio frequency component 1650–1750 Hz 1950–2050 Hz 2250–2350 Hz 2550–2650 Hz	$\leq$400 mA amplitude
High frequency component 5050–5100 Hz	$\leq$270 mA amplitude

varied in direct proportion to the speed. The power fed back to the source is proportional to the speed, decreasing linearly with speed.

(b) Mode 2 – In this region, the motor current is held constant as slip frequency is held constant. Motor voltage is adjusted as speed (frequency) falls to maintain a constant flux density in the motor. Motor voltage therefore reduces proportionally with speed (frequency). At very low speeds the electric braking is allowed to fade.

5.3   *Requirements of ac motor traction drives*

  (i) The source side converter should ensure a power factor as near unity as possible.
 (ii) The converter should be designed such that the psophometric disturbance current is kept within an enforced limit (table 1).
(iii) The converter should be designed such that no intolerable level of interference is caused to track circuit, signal or telecommunication equipment.
 (iv) The power converter/inverter should ensure four-quadrant operation, and regenerative braking should be available from maximum speed range up to standstill.
  (v) The power converter/inverter should ensure the full utilization of the available adhesion.
(vii) The inverter should incorporate beatless control system to suppress beat phenomenon.

6.   **Load-commutated inverter synchronous motor traction drives**

6.1   *Basic principle* (Dubey 1989, p. 416)

Generally a self-controlled synchronous motor drive is used for electric traction drives
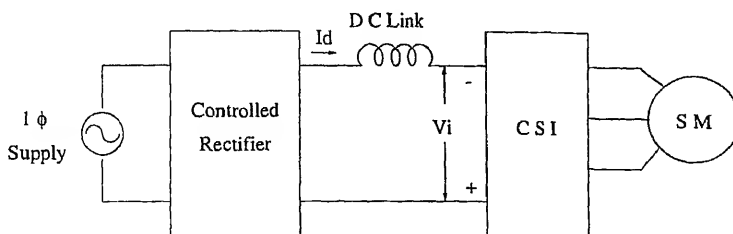
Figure 3. Load-commutated inverter synchronous motor drive.

In a self-controlled synchronous motor drive, the following control strategies are generally used.

(i) Constant margin angle control;

(ii) constant lead angle control;

(iii) constant no-load angle control.

In all the above control strategies, the synchronous motor operates with a leading power factor and the thyristors of load-side converters are commutated by the motor-induced voltages, in the same way as the thyristors of line-commutated converters. Because the firing is synchronized with the machine-induced voltage, the machine always operates in self-control mode. The supply to source side converters may be *dc* or *ac*.

A load-commutated inverter synchronous motor traction drive using thyristors is shown in figure 4. Here the supply is *dc*. The *dc* current source is formed by a two-quadrant chopper (quadrants I & VI), a *dc* link inductor and a current feedback. The load-side converter is commutated by the back *emf* of the synchronous machine (Pearson & Sen 1984).

When the overhead catenary is *ac*, the scheme shown in figure 5 is used. Here the current source is formed by source side converter (a controlled rectifier), *dc* link inductor
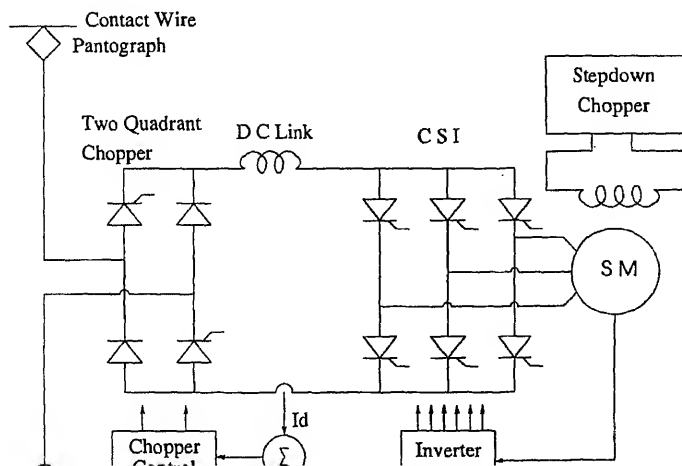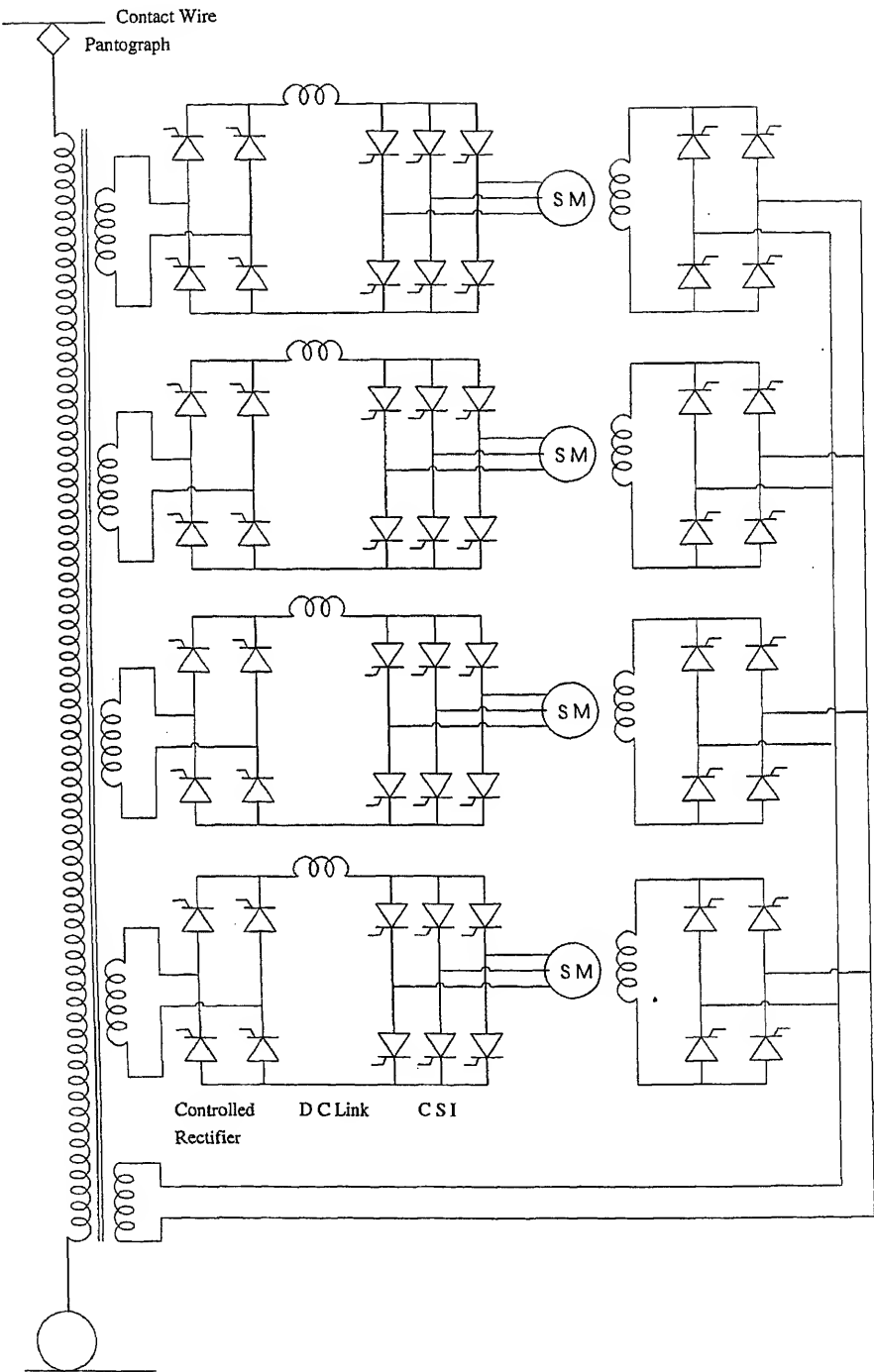
Figure 5. Regenerative synchronous motor *ac* traction drive.
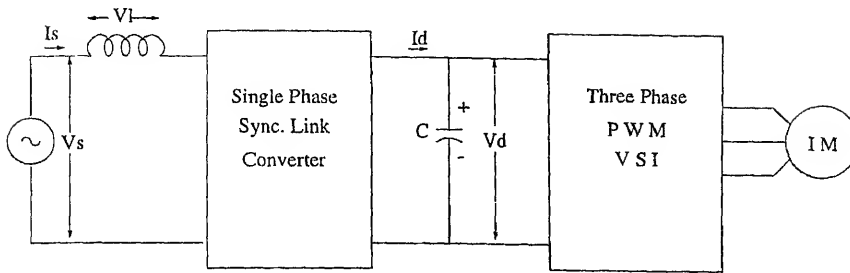
Figure 6. Regenerative induction motor drive using synchronous link converter.

d the current loop. The load-side converter is operated in self-control mode and with
ad commutation.

2 *Control of load commutated inverter synchronous motor drive*

hen the drive is in motoring, the source-side converter acts as a rectifier and the load-side
nverter acts as an inverter, resulting in the forward motoring operation. In this mode of
eration the average value of V_i (figure 3) is negative, and I_d is positive and hence the
wer flows from the dc link to the machine giving operation in quadrant I.

For braking the drive, the load side converter is controlled such that the average V_i is
sitive. Since the direction of I_d remains unchanged, the power flows from the motor to
e dc link. In this mode of operation, the load-side converter works as a rectifier and the
achine regenerates. The source-side converter is operated to feed the energy back to the
urce.

For reversing the direction of the drive, the phase sequence of the load-side converter
reversed by interchanging the control signals between switches of any two legs of the
verter.

In the case of the load-commutated synchronous motor drive, at low speeds the induced
nf will be insufficient to commutate the thyristors of the load-side converter. Therefore,
start and for speeds below 10% of base speed, the commutation of the load-side converter
yristor is done by forcing the current of conducting thyristors to zero, with the help of
e source-side converter (Peterson & Frank 1972).

Alternatively, a synchronous motor can be started by employing a simple forced-
mmutation circuit utilizing a single commutation capacitor and two auxiliary thyris-
rs for the entire inverter (Steigerwald & Lipo 1979).

VSI-squirrel cage induction motor drive

1 *Basic principle*

regenerative induction motor drive using a synchronous link converter is shown in
gure 6. The synchronous link converter permits the realization of an economic regenera-
ve ac drive with a power factor of nearly unity and a low harmonic current in the source
rrent. The converter circuit is shown in figure 7.
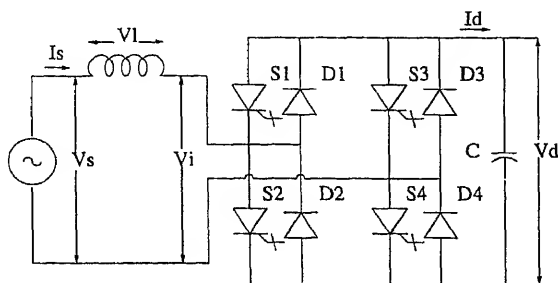
Figure 7. Single phase synchronous link converter.

When the single-phase voltage source inverter is operated with pulse-width modulation, it produces the voltage $\bar{V}_i$. If the voltage V_d is maintained constant, the magnitude of V_i may be varied by adjustment of the modulation index of the converters. The phase of $\bar{V}_i$ w.r.t. $\bar{V}_s$ can be altered by changing the phase of firing pulses of the inverter, w.r.t. the phase of $\bar{V}_s$. By appropriately choosing the magnitude and phase of $\bar{V}_i$ for a given $\bar{I}_s$, the phase angle between $\bar{V}_s$ and $\bar{I}_s$ can be adjusted to a desired value.

Figure 8 shows the phasor relationship of various vectors when the phase angles between the source voltage $\bar{V}_s$ and source current $\bar{I}_s$ are $0°$ and $180°$ respectively. In either case the power factor is unity. When the phase angle is zero, the converter works as a rectifier transferring power from ac source to the dc link. In this mode of operation, current I_d has a positive direction. When the phase angle is $180°$ current I_d reverses and the power flows from the dc link to the source and the converter works as an inverter. Because of the operation of the converter with pulsewidth modulation, the source current has a low harmonic content.

Figure 9 shows a three-phase regenerative traction drive using a synchronous link converter and PWM inverter. Here two synchronous link converters are connected in parallel feeding an inverter, which in turn feeds three induction motors connected in parallel. Two such units are fed from a common transformer.

Depending upon the power level IGBTs/GTOs can be used in the construction of a synchronous link converter. In case of Electrical Multiple Units (EMU), IGBTs can be used both in synchronous link converter and inverter. Since the IGBTs can be operated at a frequency of 2.5 kHz the harmonics in the source current can be reduced to a smaller value. We can use sinusoidal PWM, current controlled PWM or vector PWM for both converter and inverter. In case of a locomotive, because of the large power levels, GTOs have to be used. Since at high power levels, GTOs are operated around 400 Hz, generally sinusoidal PWM is used and more than one converter is operated in parallel to keep the harmonics in the source current within acceptable limits.
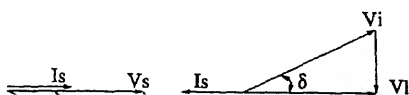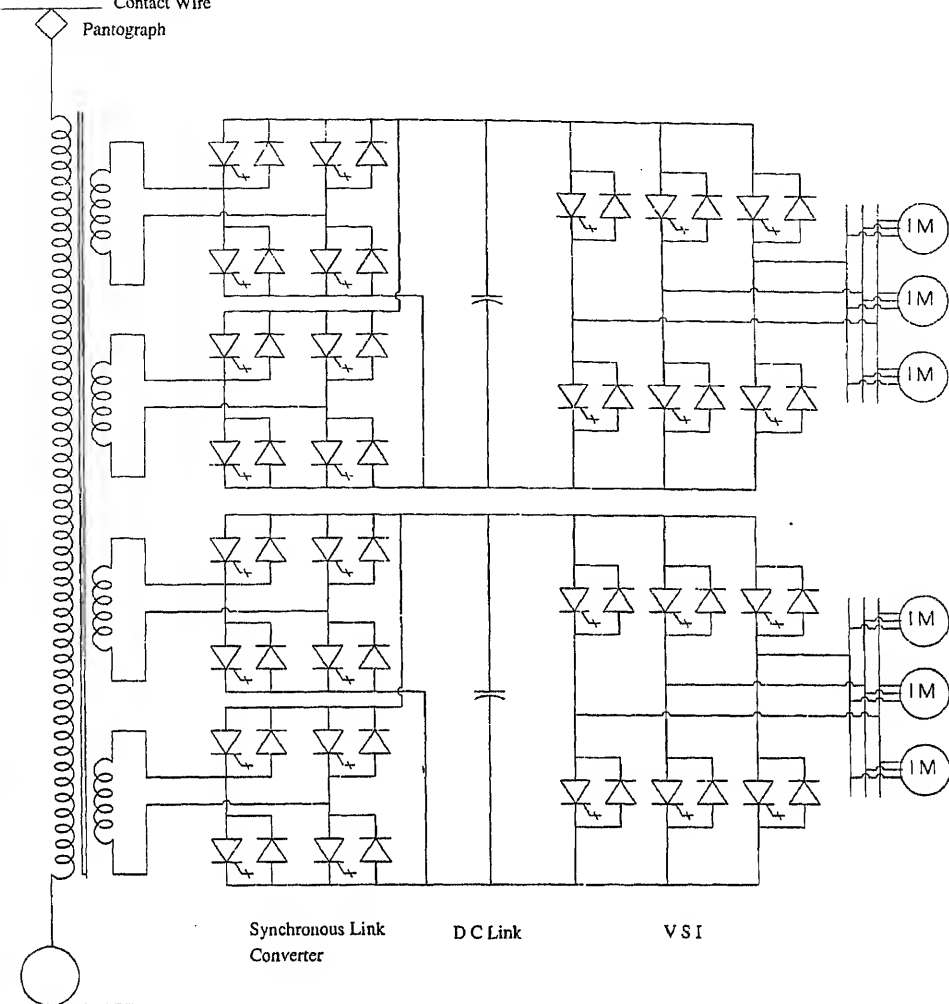
Figure 9. Regenerative traction drive using synchronous link converter and PWM VSI.

7.2 Control of VSI – Squirrel-cage induction motor traction drive

The modern VSI squirrel-cage induction motor traction drives are generally torque-controlled drives. The drive is started with reduced voltage and frequency. This frequency reduction improves the motor power factor and increases the torque per ampere at the start. Hence the rated torque is available at the start, and the drive is accelerated to its operating speed. The drive is operated as described in § 5.2. Here the power is transferred from the source to the load.

For transferring the operation from motoring to regenerative braking, the inverter frequency is reduced. A reduction of inverter frequency makes the synchronous speed less than the motor speed and transfers operation from quadrant I to quadrant II. Here the power flows from motor to *dc* link and from *dc* link back to source. The inverter voltage and frequency are reduced to brake the machine to zero speed.
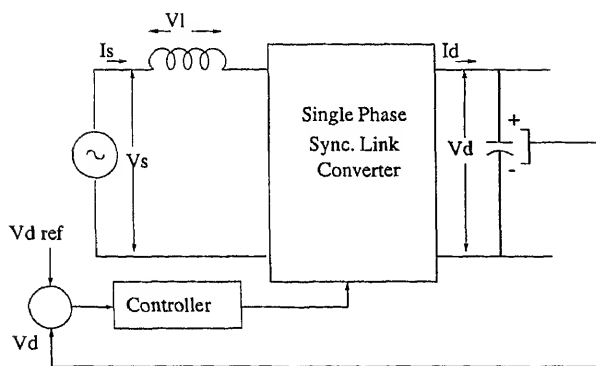
Figure 10. Closed loop control around a synchronous link converter.

For speed reversal, the phase sequence of the inverter voltage is reversed by interchanging the control signals between the switches of any two legs of the inverter. In the scheme described above, when the source is unable to take back the regenerated power, a dynamic braking scheme can be used.

For this a braking resistor is connected in series with a switch across the *dc* link capacitor 'C'. The generated power charges the filter capacitor and its voltage rises. When the filter capacitor voltage reaches a prescribed maximum value, the switch is turned on, connecting the braking resistor across the capacitor. The energy generated and the energy supplied by the filter capacitor are dissipated in the braking resistor. The capacitor voltage falls. When the capacitor voltage reaches a prescribed minimum value the switch is turned off. Again the capacitor voltage starts rising and the cycle repeats. Thus only that portion of the regenerated energy is dissipated by dynamic braking which cannot be accepted by the source.

The following control techniques are generally used with synchronous link converter.
(1) Current-controlled PWM;

(2) Indirect current control (sinusoidal pulsewidth modulation and its modifications).

With the above mentioned current controlled techniques, for a given value of reference, the power input to the converter has a fixed value. When the load on the converter is decreased, there is imbalance between the power input and the power output. Hence a closed loop voltage control is provided across the synchronous link converter as shown in figure 10.

8. Harmonic reduction and control of torque pulsation

The line side converter generates current harmonics which are injected into the overhead lines. Hence, measures are taken to keep the harmonics below a desired limit. Some of the methods used with VSI squirrel cage induction motor drive are as below.
  (i) By using series/parallel connection of converters, the supply current harmonics can be reduced considerably. If N converters are put in series/parallel and if the carrier signals of individual converters are equally displaced by an angle π/N, then some of the reflected harmonics cancel on the primary side of the transformer. The ripple
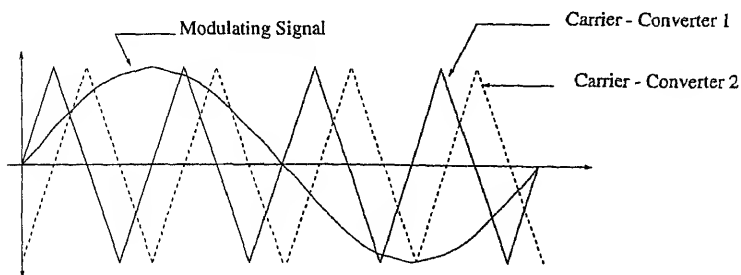
Figure 11.   Carrier signals for two converters operating in parallel.

frequency of the primary current has a frequency $(N \times fc)$, where N is the number of converter put in parallel/series and fc is the carrier signal frequency.

For example, if the synchronous link converter is realized by using two GTO converters, say converter I and converter II in parallel, then the firing of the GTOs in converter I and converter II is done by using a common modulating signal and the triangular carrier wave phase shifted from each by an angle 90° (because $N = 2$) as shown in figure 11.

) By using three-level converters as compared to two-level converters.

) By employing switched electronic compensator which generates an exact replica of the harmonic current. It is fed into an extra winding of traction transformer to produce harmonic counter emf. This ensures that the transformer main flux is sinusoidal, and hence so is the induced voltage in the primary. The line current is low-pass filtered by the transformer leakage inductance and assumes a sinusoidal waveform.

For reducing the harmonics in the inverter output voltage and hence to minimize the rque pulsations, the inverters can be connected in parallel. They share the common dc k and the ac output terminals of the respective phases are connected through an inter-ase reactor. The mid-point of the inter-phase reactor is connected to the respective phases an induction motor. The inverters are operated with common modulating signals, but th the carrier signals of individual inverters being equally displaced within the carrier riod interval.

Advantages of VSI squirrel-cage induction motor traction drive

e VSI squirrel-cage induction motor traction drive has the following advantages.

(i) Power factor of unity.

(ii) Low harmonic content in the source current.

iii) Regenerative braking capability at low cost.

iv) Smooth acceleration due to absence of low speed torque pulsations.

(v) Efficient and quiet operation due to low harmonic content in the motor input voltage.

vi) Good adhesion due to fast dynamic response and absence of torque pulsations.

vii) Multimotor drive possible due to use of VSI.

(viii) Regenerative braking capability offering energy saving.

(ix) Dynamic braking can be resorted in the event of catenary failure due to use of VSI.

(x) Improved energy efficiency.

(xi) Reduced maintenance cost.

(xii) Reduction in the size and weight of traction circuit equipment.

(xiii) Possibility of reducing the capacity of the traction substation, because the power factor of the system is always unity, both in powering and braking.

10.   Conclusions

AC motor traction drives are superior to DC motor traction drives. Because of recent advances in power electronics, drive control techniques and availability of micro-controllers for control and fault diagnosis have enabled the realization of traction drives with high efficiencies, high levels of performance and low down-times.

References

Dubey G K 1989 *Power semiconductor controlled drives* (Englewood Cliffs, NJ: Prentice Hall)

Dubey G K 1993a Semiconductor controlled AC motor traction drives. Course Package, Continuing Education Program of Indian Society for Technical Education

Dubey G K 1993b Semiconductor controlled DC motor traction drives. Course Package, Continuing Education Program of Indian Society for Technical Education

Dubey G K 1994 *Fundamentals of electric drives* (New Delhi: Narosa)

Dubey G K 1997 A modern 25 kV squirrel cage induction motor drive for traction. *National Conference on Electric Drives and Control for Transport Systems*, Vidisha

Dubey G K, Rao K C (eds) 1993 *Power electronics and drives* (New Delhi: Tata McGraw-Hill) vol. 1

Dubey G K, Doradla S R, Joshi A, Sinha R M K 1986 *Thyristorised power controllers* (New Delhi: Wiley Eastern)

Holtz J, Onnokrah J 1992 Adaptive optimal pulsewidth modulation for the line side converter of electric locomotive. *IEEE Trans. Power Electron.* PE-7: 205–211

Kanetkar V R, Dubey G K 1995 Bidirectional AC to DC power conversion – Trends, developments and application of concern. *Indian Elec. Eng. Manuf. Assoc. J.*: 7–28

Kielgas H, Nill R 1980 Converter propulsion system with three-phase induction motors for electric traction vehicle. *IEEE Trans. Ind. Appl.* 16: 222–223

Kliman G B 1972 Harmonic effects in pulsewidth modulated inverter induction motor drive. *IEEE Ind. Appl. Soc. Annu. Meeting* pp 783–790

Lienau W, Muller Hellman A, Skudelny H C 1980 Power converters feeding asynchronous traction motors of single phase AC vehicles. *IEEE Trans. Ind. Appl.* 16: 103–110

Nakamura K, Kimura A, Iwataki M 1988 Inverter drive system for AC electric rolling stock. *Hitachi Rev.* 37: 71–376

rson, Sen P C 1984 Brushless DC motor propulsion using synchronous motors for transit systems. *IEEE Trans. Ind. Electron.* IE-31: 346-351

erson T, Frank K 1972 Starting of large synchronous motor using static frequency converter. *IEEE Power Appl. Syst.* 91: 172–179

nkett B, Plette D I 1977 Inverter induction motor drive for transit car. *IEEE Trans. Ind. Appl.* 13: 26–37

bertson S D T, Hebbar K M 1971 Torque pulsation in induction motor with inverter drivers. *IEEE Trans. Ind. General Appl.* 7: 318–323

eigerwald R L, Lipo T A 1979 Analysis of a novel forced commutation starting scheme for a load commutated synchronous motor drive. *IEEE Trans. Ind. Appl.* IA-15: 14–24

kegawa T, Kamiyama K, Takahashi J, Kimi, Matsutake M 1992 A multiple PWM GTO line side converter for unity power factor and reduced harmonics. *IEEE Trans. Ind. Appl.* IA-28: 1302–1308

sunami M, Nakamura K, Sakane T 1986 Application of power electronics to rolling stock. *Hitachi Rev.* 6: 305–310

Subject Index

Author Index